Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

Dijkshoorn, C.R.

2019

document version
Publisher's PDF, also known as Version of record

Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

Chris Dijkshoorn

Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

Chris Dijkshoorn

Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

Chris Dijkshoorn

SIKS Dissertation Series No. 2019-06

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

COMMIT/

The research reported in this thesis was funded by the Dutch national program COMMIT/.

Printing Ridderprint BV | www.ridderprint.nl

VRIJE UNIVERSITEIT

Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op woensdag 3 april 2019 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Chris Ridder Dijkshoorn

geboren te Lelystad

promotoren:	prof.dr. A.Th. Schreiber
	prof.dr. L.M. Aroyo
copromotor:	dr. V. de Boer

promotiecommissie: prof.dr. F.A.H. van Harmelen
prof.dr.ir. G.J.P.M. Houben
prof.dr. I.B. Leemans
dr. J.I.E.M. Mäkelä
prof.dr.ir. R. Verborgh

ACKNOWLEDGEMENTS

It feels strange, the first (and maybe only) part that you read of this thesis is the last part I write. Writing this thesis has been a lengthy and sometimes challenging process, during which many people supported me. Working together with so many fabulous people made this process worth the while and allowed me to see it through to the end.

First of all, I would like to thank my promotors. Guus for his calm, always able to get the bigger picture in a snap. His insights in the domain helped guide me to get the contents of this thesis into a coherent whole. The knowledge he displayed about modeling data, challenged me and sparked my interest in such a way, I ended up doing it for a living. I thank Lora for her endless stream of interesting ideas and suggestions. In the beginning, I had to learn how to cope with them and to not try to pursue them all at once. Cherrypicking as an art. At the same time, she taught me how to do many things at once, while keeping my goals in sight. This efficiency is a skill useful for the rest of my life. Victor became my co-promoter in the final stages of my Ph.D. His positive attitude and prompt comments helped improve texts quickly and made the writing process joyful. This, in combination with a laid-back attitude, led me to believe that it would be possible to actually finish my thesis.

Many others helped, albeit not in an official capacity. Jacco co-authored many of the papers included in this thesis, nudging them into proper scientific texts. His pace of thinking was so fast, I often could not keep up, eyes glazing over during meetings. I hope this was not too apparent. Many times I walked towards Jans office, with programming problems on my mind. He taught me how to formulate questions in such a way, that in the end, I could turn around half-way the corridor since I was able to answer the question myself. I thank Stefan for all his help and especially for his green elephant drawing skills. I want to thank Rinke for his help, in a period I could really use it. By sharing his Ph.D. experience, he helped me understand that sometimes you need to deviate from the beaten track.

Over the years I got to enjoy the company of many office mates. It was nice to have such a composed person as Archana in the office. I am sorry for all the times I made her jump when I cursed some line of code. It was an honor being her paranimf and I hope she enjoys being mine. I started my Ph.D. at almost the same time as Valentina, working on a similar topic. It is a shame we went our own ways and did not manage to do more research together. It was really nice to share an office with her, though, and to see her master students whither

when they did not do what she wanted. It was great to work with Cristina on multiple projects. It has been a pleasure to see her grow over the years. Martine is a biologist by heart. It was interesting to be part of her journey into the ICT world. I hope she continues doing what she loves and that it will contribute to a better world. I met Sara in Como in Italy, sometime before starting my Ph.D. She is one of the brightest people I know and I am happy she endured the Dutch food to finish her Ph.D. in Amsterdam. Laurens has technically never been an office mate, but the number of times he burst into my office to ask whether to go for a coffee, can amount for the time normal people spent around me in the office. He taught me that playing a game can be enjoyed while you do not win, although he probably will not agree with me on that.

Taking a long time to finish your Ph.D. also means that you see a lot of people come and go in a research group. I will probably not get around mentioning everyone, but please do appreciate the effort. I envy how Tobias is able to structure his day, while supervising more students than I can count. In Davide we trust. At times, the enthusiasm of Willem was almost intimidating, although I also believe that made him a great driving force of the group. Michiel was always able to hack something nice together and it comes as no surprise that he now works for a company with 'an app in a day' as a motto. I enjoyed discussing Iran with Dena and greatly respect her scientific attitude. The discussions with Niels about how to turn the things we build into something scientific, were always constructive. I am grateful for the CrowdTruth team, consisting out of Anca, Oana and Benjamin, for always having my back on crowdsourcing related matters. It was more than interesting to get to know (a bit) about the different fields of expertise of Jan, Ronald, Chris, Antske, Paul and Laura. Having so many Ph.D. students around to go for a coffee or drink always made my day. Thanks for that Xander, Astrid, Riste, Jesper and Anna. I would also like to acknowledge Mojca, Caroline and Elly for keeping this bunch of nutty scientists in line.

Marieke was the first colleague with whom I co-supervised a master student. Maybe not the greatest idea to start supervising students in my first year, but nonetheless, I really enjoyed it and learned a lot. After Guido, more students followed: Jorik, Dick, Quinten and Kieke. While some of their work contributed to this thesis, I hope I also gave them enough freedom to pursue their own interests. Supervising students in interdisciplinary Network Institute projects was always a joy, because you learn so much from each other. The INVENiT projects, the first involving Wesley en Thijs, were co-supervised by Inger. From Inger, I learned a lot about how things are done in the humanities. Cristina and Leon were part of the second INVENiT project. It was incredible to see how open Leon was to get to know computer science methods and it is great to know both of them started their own Ph.D.

The close collaboration on Semantic Web related topics with the Knowledge Representation and Reasoning group was inspiring. Together, Annette, Frank and Stefan make a strong foundation for a group that provides good education, something which I enjoyed first-hand during my study Artificial Intelligence. I always admired Frank's ability to ask the questions, that allows one to tell a better story. It was nice to be able to join lunches (and the occasional poker game) with researchers like Albert, Ali, Filip, Finn, Kathrin, Krystyna, Marat, Peter, Veruska and Wouter. Our groups also hosted many inspiring visiting researchers such as Silvia, Liliana and Joe. Furthermore, I enjoyed many SIKS courses and met some wonderful Ph.D. candidates there, like Michiel, Maya and Hugo.

Within the SEALINCMedia project, I collaborated with some great researchers. Myriam and Mieke were the dynamic duo from CWI. I thank Mieke for the many funny highlights of differences between the Flemish and Dutch. Myriam always had a constructive attitude, resulting in helpful discussions about doing proper research. It was nice to visit Delft and work together with Jasper on crowdsourcing setups, learning a thing or two about software engineering. Wan, Geert-Jan and Alessandro, were always willing to help provide a theoretical foundation for what we were building. Antoine and Hugo represented Europeana and showed how our research could be translated into useful insights for industry.

Some have already been mentioned above, but I want to sincerely thank my examination committee. Thank you Frank van Harmelen, Geert-Jan Houben, Inger Leemans, Eetu Mäkelä and Ruben Verborgh for taking the time to read this thesis, and for approving it.

Since it was the main use case of my project, I had the pleasure to work with many people at the Rijksmuseum. Lizzy was great to have as a partner on the project, always able to support us with data and incredible stories. Henrike gave us upfront and sincere feedback, which was way more constructive than 'your system looks pretty'. From Inge, I learned quite something about data cleaning. Trineke was an incredible source of insights in the workings of systems, backed by a throve of art-historical knowledge. The *vier-uur-momentjes* of Bas, Caroline, David and Lotte were a fun way to get my head out of programming mode. Thanks Esther, Marcel, Ralph and Inge for coping with my complaints about finalizing a thesis. Unfortunately, some research takes a bit longer to finish. I thank Saskia for her support and patience. In general, I want to thank everyone in *toren 5* for being so welcoming and supportive. I hope to work with all of you for many years to come.

The crowdsourcing campaigns would not have been possible with the help of many others. Maarten served as a contact with Naturalis and his enthusiasm for our work was flattering. I am really happy he joined the ranks of the Rijksmuseum. Let us see when we will run

another campaign. Sebastien of the library of the Vrije Universiteit helped us mobilize bible enthusiasts. His curiosity into the world of computer science hopefully made it an interesting exercise for both sides. Lisa and Ykje of Modemuze enabled us to address fashion experts. Of course, the main drivers of these campaigns were all the participating enthusiasts. It made this research so much more interesting, everybody who so eagerly shared their knowledge and experience!

My research visit to the British Museum was an incredible experience. I loved the discussions with Dominic about data modeling and grew into an enthusiast for everything reeking of CIDOC-CRM. Barry kept the ResearchSpace team grounded and I greatly respect his ability to translate theoretical models into usable programs. Sarah was a great host and Alan showed me there is art in interfaces. I stayed with a lovely couple of anthropologists, who managed to make me go vegan for some months. Despite that, it was a joy to meet so many interesting people at Dundalk Road.

Like many of the projects mentioned above, DigiBird was a multidisciplinary effort. Johan and Maarten at the Sound & Vision institute (which I probably miswrote for the umpteenth time) kindly hosted us for six months. I really appreciated the practical approach and involvement of Maarten in the project. Oftentimes Jaap helped us out with technical support. Sander from Naturalis made a lot happen, including finally some real use of something we developed.

I became friends with Bas, Kees and Koen during my studies. It is intriguing to see how different the things we ended up doing are, which makes our occasional meetings all the more interesting. While Bas and I were writing our master thesis in Como, Emanuele and Spyros showed me that doing research can be enjoyable (contrary to my belief back then). I thank Helen for sharing her Ph.D. experience from a biologists point of view, it was wonderful to visit your 'office' in Panama. The friendship of Jelle, Joost, Justin, Maarten, Mieke, Remco, Tristan and Wouter goes back to my high school period. They have always been there to support me. Remco and Mieke take up a special spot in that list, since they ended up being my roommates for multiple years. Sorry for all the nights of working and being miserable after that.

My parents and sister have had a lot to endure during my educational career. Ranging from an extremely lazy adolescent, to a disillusioned wannabe architect, to a somewhat disciplined artificial intelligence student, to a researcher on the verge of being a workaholic: they were there with advice when I needed it the most. I know my father is still a bit surprised, but extremely happy, that it worked out well in the end. I was also lucky to have a family in law that was familiar with and supportive of Ph.D. students. But most of all I want to thank Merel. She stood by me, no matter how difficult things got.

CONTENTS

1

INTRODUCTION

The Semantic Web is a web of interoperable data, that can be interpreted both by humans and machines. Cultural heritage institutions recognized the value of Semantic Web technologies and a lot of work has gone into publishing collections, creating standardized data models and compiling structured vocabularies. In this thesis, we analyze Linked Data published by cultural heritage institutions and investigate how such data can be further contextualized and enriched. We show how enrichments improve access to online collections, as well as support the integration of collections of different institutions. We start this introduction with background information about online cultural heritage collections, nichesourcing and the Semantic Web.

1.1 BACKGROUND

1.1.1 *Cultural heritage collections online*

Many cultural heritage institutions provide online access to digital representations of their collection objects[1]. Object retrieval and discovery methods, that allow various users effective access to collections, rely on data that describes these objects (i.e. metadata). Examples of metadata are the creator, material and subject matter of an object. Most cultural heritage institutes already have basic metadata at their disposal, for collection management purposes.

Museums should be hesitant to publish metadata online that comes straight out of the collection management system, as it creates two major problems. First, online visitors are deprived of the narrative and context provided by the meticulously composed exhibitions in museums. Second, the well-curated object descriptions in those exhibitions are missing online, since objects in collection management systems are typically described with minimal information, in art-historian jargon and for the purposes of preservation [32]. As a result, despite the fact that numerous museums have already published their collection online, it remains a challenge to explore them. To address this problem, we investigate how existing metadata can be contextualized using annotations.

1 For example:
`http://www.metmuseum.org/collections`
`http://www.britishmuseum.org/collection`
`http://www.louvre.fr/moteur-de-recherche-oeuvres`

1.1.2 *Nichesourcing annotations*

To improve access to online collections, cultural heritage institutions started initiatives to better describe objects by gathering annotations. However, the size of many collections, as well as the diversity of the topics covered, goes beyond any number that in-house annotators can handle in a feasible amount of time and with the desired level of quality. An opportunity to address this problem presented itself with the rise of crowdsourcing [24]. Many museums embraced the crowd to annotate collection objects and to bridge the gap between professional descriptions and user expressions [57]. Some well-known examples are the Steve.museum and Your Paintings projects [36, 74]. Crowdsourcing not only addressed the annotation problem but also created new ways of engaging audiences online [64].

The crowd proved to be a quick and inexpensive source of large quantities of annotations, but not enough mechanisms were in place to ensure the quality of added information, especially for knowledge-intensive annotation tasks, such as often encountered in the cultural heritage domain. Therefore, we introduced *nichesourcing*, an extension of the crowdsourcing paradigm, aimed at solving knowledge-intensive tasks, by identifying and engaging small groups of amateur experts, rather than addressing the "faceless" crowd [88]. A niche is gathered from either distributed experts on a specific topic or from an existing network centered around the same culture, location or topic. In both cases, the members possess domain knowledge and are intrinsically motivated to contribute and provide high-quality results. In Chapter 4 of this thesis we introduce the Accurator nichesourcing methodology, detailing how nichesourcing can be used to gather high-quality annotations for different types of collections.

1.1.3 *The Semantic Web*

Semantic Web technologies allow cultural heritage institutions to publish interconnected, interoperable data, with explicit semantics. Currently, most information on the internet is published as unstructured text. Unstructured text is hard to interpret by machines, which makes automated reasoning over statements and integration of information difficult. The Semantic Web transforms the internet from a web of documents into a web of data [6]. Many cultural heritage institutions have started to publish data on the web [43, 50, 72, 90].

Statements are made in the form of triples, consisting of a subject, predicate and object. An example of such a triple is: *:nightwatch dc:creator :rembrandt*. Here, the *dc* prefix indicates that the predicate "creator" originates from the standardized Dublin Core data model. The subject and predicate are always represented by an identifier, whereas the object can be an identifier as well as a literal. To illus-

trate, for the object of this triple, we could also have used the literal *"Rembrandt van Rijn"*. Unlike literals, identifiers allow for disambiguation of resources. If, for example, two artists have the same name, an identifier helps to distinguish between them. In addition, using the same identifiers for resources in different datasets establishes a common ground, aligning information published by different institutions. We use this principle in Chapter 6, to integrate collections of different modalities.

Statements can also be used to make the semantics of data explicit. Collections of such statements forming a formal, explicit and shared conceptualization are called ontologies, which are instrumental for correctly interpreting data [71]. The explicit semantics and use of identifiers eases reuse and integration of data. This aligns with the goal of cultural heritage institutions to disseminate information, but also provides the opportunity to reuse complementary information published by other institutions. Besides using ontologies to structure information about collection objects, structured vocabularies can be used to describe objects with a normalized set of concepts, that can be reused by multiple institutions. These structured vocabularies are an ideal source of concepts for annotating collection objects within crowdsourcing initiatives.

1.2 PROJECT CONTEXT

The research reported upon in this thesis is conducted in the context of the project Socially-Enriched Access to Linked Cultural Media (SEALINCMedia)[2], part of the Dutch national program COMMIT/[3]. The goal of the SEALINCMedia project is to improve access to multimedia collections, by using crowdsourcing and automated methods to enrich content. The project focusses on two application domains, automated video analysis and crowdsourced enrichments of cultural heritage collections. Earlier work of the MultimediaN project served as a foundation for the research in this thesis, providing collection alignments, a graph search algorithm and an annotation interface [67].

The application domain of this thesis is cultural heritage, a domain in which providing access to online collections remains a challenge. We closely collaborated with the Rijksmuseum Amsterdam[4], which gave us access to collection data and search logs from the website. The museum hosted two annotation events, during which we could collect information about how experts enrich collection data. A similar event was held at the University Library of the Vrije Universiteit Ams-

2 http://sealincmedia.wordpress.com
3 http://www.commit-nl.nl
4 http://www.rijksmuseum.nl

terdam[5], with which we collaborated in the context of the Network Institute project INVENiT[6]. A research visit to the ResearchSpace team[7] at the British Museum[8] provided valuable insights into cultural heritage data modeling practices.

The COMMIT/ program granted funding for valorizing research results in the DigiBird project[9]. This project brought together four institutions with different, but complementary collections. In addition to the Rijksmuseum, the Naturalis Biodiversity Center[10], the Xeno-Canto foundation[11] and the Netherlands Institute for Sound and Vision[12] were part of this assembly. All four institutions provided access to collection data, while the Sound and Vision institute kindly hosted two researchers and provided resources for the required infrastructure. The Naturalis Biodiversity Center incorporated the results of the DigiBird project into their Dutch Species register website[13].

1.3 RESEARCH QUESTIONS

The main objective of this research is to formulate *a reusable method to gather annotations that enrich and contextualize objects, thereby improving access to online cultural heritage collections*. We guide our research through answering four research questions: two regard contextualization and enrichment, the other two questions regard access to online collections. Through the first research question, we investigate how existing ontologies cope with the challenges of capturing data about cultural heritage objects and their context.

> *1. How do different modeling approaches influence the contextualization of cultural heritage collections published online?*

Cultural heritage institutions possess a wealth of knowledge about their collection objects. It is not trivial to translate this knowledge into data that is fit to be published online. We investigate how existing ontologies cope with modeling challenges encountered in the cultural heritage domain, as well as their impact on publishing contextual data about collection objects. The second question regards how additional information can be added to cultural heritage objects by involving niche communities.

> *2. What method for engaging niche communities to enrich cultural heritage objects can result in high-quality annotations?*

5 http://www.ub.vu.nl
6 http://invenit.wmprojects.nl
7 http://www.researchspace.org
8 http://www.britishmuseum.org
9 http://www.digibird.org
10 http://www.naturalis.nl
11 http://www.xeno-canto.org
12 http://www.beeldengeluid.nl
13 http://www.nederlandsesoorten.nl

Institutions often have basic metadata about objects for collection management purposes. This data is not always sufficient for providing access to collections online. Extending this information requires specialized knowledge about an object and its subject matter. We investigate whether it is possible to formulate a crowdsourcing methodology that involves niche communities in the annotation process of collection objects. Niche communities are targeted since it is of paramount importance for cultural heritage institutions that enrichments are of high quality. The third question investigates how these enrichments can be used to generate more diverse search results.

> *3. How do enrichments from various structured vocabularies influence the diversity of search results?*

Cultural heritage collections contain many objects, of which the details are often unknown to the average user. Presenting more diverse search results can help users to explore collections and reach more relevant objects related to their search query. A dense web of contextual information can support explorative search. Previous questions explored adding context and enrichments to objects, here we investigate how contextual information from different structured vocabularies impacts the ability to retrieve diverse search results. The fourth question also regards access but focusses on how different cultural heritage collections can be integrated online.

> *4. How to address the issue of continuously evolving data in the process of integrating cultural heritage datasets from various institutions?*

Collections from different institutions can be complementary. For example, an oeuvre of an artist is rarely kept at one institution and a certain subject matter topic can be highlighted with media from different modalities. A more complete picture can be given to users when it is possible to provide integrated access to multiple collections at once. Here, we focus on how this integration is possible for datasets that are subject to constant change, due to for example a continuous stream of added enrichments.

1.4 APPROACH

The five-step method presented in this thesis contextualizes objects, in order to improve access to online cultural heritage collections. Each step of the method provides input for the next step, as illustrated by Figure 1. To lay a foundation for analysis of contextualization and enrichment of data, we first describe a cultural heritage collection published online, originating out of a real-world setting. To do so, we analyze Linked Data published by the Rijksmuseum. In collaboration

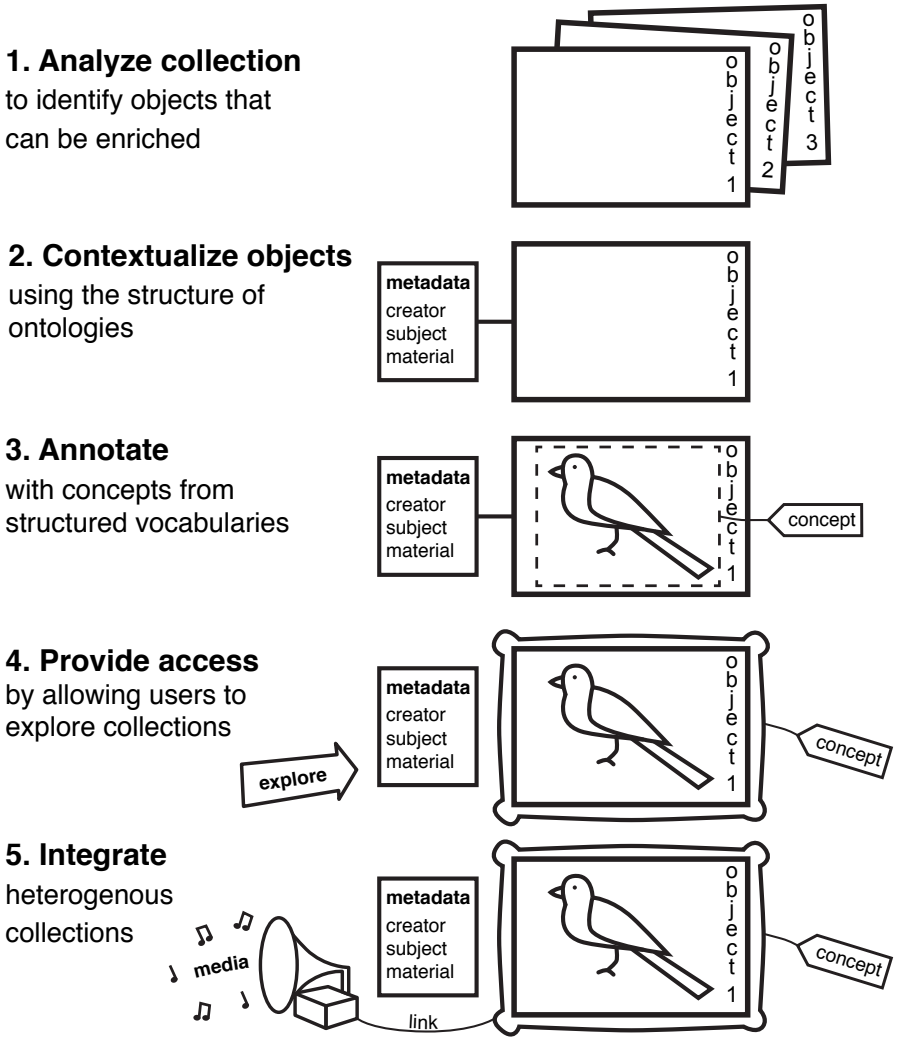

Figure 1: The five-step method of this thesis.

with the museum, we describe the data conversion process, gather statistics about the published metadata and create an overview of the links to external data sources.

The results of data analysis and observed data publishing challenges serve as input for a comparison of cultural heritage data modeling approaches. To answer the *first research question*, we compare common modeling approaches, to see how ontologies support the contextualization of cultural heritage objects published online. We discuss how modeling approaches of two commonly used cultural heritage ontologies address these challenges, illustrated with examples from the Rijksmuseum collection.

After analyzing the data and ontologies, we continue with the topic of enrichment. Crowdsourcing is potentially a source of numerous annotations but suffers from drawbacks when poorly implemented. For answering the *second research question*, we develop a nichesourcing methodology, which is repeatable and assures that high-quality

annotations are gathered. To evaluate the method we run three experiments in the form of nichesourcing campaigns, during which niche-experts contribute their knowledge, by extending the descriptions of cultural heritage objects. Participants evaluate the annotation tool and nichesourcing method by filling in a questionnaire. The collected annotations are compared to a gold standard when available or are reviewed by professionals.

We continue by investigating how access to online collections is improved by the contextualization and enrichment of objects. To measure the impact of enrichments on search result diversity we run an experiment with search queries obtained from the logs of the Rijksmuseum website as input. These queries are used in combination with an existing graph search algorithm, which leverages the connections between concepts and collection objects, to obtain relevant results. In order to answer the *third research question*, we alternate different enrichments from various structured vocabularies. We measure how the different characteristics of the alignments between collection and vocabularies and the characteristics of the vocabularies (e.g. number of links between concepts) impact the diversity of search results.

In order to answer the *fourth research question*, we develop a collection integration platform in a real-world setting, involving different methods of programmatically accessing information about objects from different cultural heritage collections. The collections are continuously extended by crowd contributions or enriched with annotations. Therefore we investigate methods that do not ingest all data once, but that integrate the latest data from different sources upon request. Thereby we compare different methods of accessing data and evaluate the benefits of the use of identifiers and standardized data models.

1.5 THESIS OUTLINE

The first part of this thesis regards enriched cultural heritage data. *Chapter 2* starts with a description of the Rijksmuseum dataset. In *Chapter 3*, we investigate modeling approaches of two cultural heritage ontologies aimed at contextualizing objects, by discussing challenges encountered while publishing data online. *Chapter 4* shifts the focus towards data enrichment. This chapter includes a detailed description of the Accurator nichesourcing methodology, a description of the Accurator annotation tool and an evaluation of these in the form of a report of three nichesourcing campaigns.

The second part of this thesis investigates how enriched data can be used to improve access to datasets. *Chapter 5* describes how various structured vocabularies, aligned with Rijksmuseum data, influence the ability to obtain diverse search results. In *Chapter 6*, we introduce the DigiBird collection integration framework, used to integrate

volatile data on the fly. *Chapter 7* includes the conclusions drawn from this thesis.

1.6 PUBLICATIONS

This introduction is based on a paper presented during a doctoral consortium and a position paper:

- Chris Dijkshoorn. "Grasping the long tail: personalized search for cultural heritage annotators." In: *Proceedings of the 21st international conference on user modeling, adaptation, and personalization*. (Rome, Italy). Ed. by Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro. Vol. 7899. UMAP '13. Springer Berlin Heidelberg, June 2013, pp. 392–395
- Victor de Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. "Nichesourcing: harnessing the power of crowds of experts." In: *Proceedings of the 18th international conference on knowledge engineering and knowledge management*. (Galway, Ireland). Ed. by Annette ten Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Acquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez. EKAW '12. Springer Berlin Heidelberg, Oct. 2012, pp. 16–20

I contributed the Rijksmuseum use case to the position paper. Publications on which the chapters in this thesis are based:

- **Chapter 2:** Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco Van Ossenbruggen, Guus Schreiber, Wesley ter Weele, and Jan Wielemaker. "The Rijksmuseum collection as Linked Data." In: *Semantic web journal* 9.2 (2018), pp. 221–230
- **Chapter 3:** Chris Dijkshoorn, Lora Aroyo, Jacco Van Ossenbruggen, and Guus Schreiber. "Modeling cultural heritage data for online publication." In: *Applied ontology* 13.4 (2018), pp. 255–271
- **Chapter 4:** Chris Dijkshoorn, Victor de Boer, Lora Aroyo, and Guus Schreiber. "Accurator: nichesourcing for cultural heritage." In: *Human computation journal* (in press)
- **Chapter 5:** Chris Dijkshoorn, Lora Aroyo, Guus Schreiber, Jan Wielemaker, and Lizzy Jongma. "Using Linked Data to diversify search results: a case study in cultural heritage." In: *Proceedings of the 19th international conference on knowledge engineering and knowledge management*. (Linköping, Sweden). Ed. by Krzysztof Janowicz, Stefan Schlobach, Patrick Lambrix, and Eero Hyvönen. EKAW '14. Cham: Springer International Publishing, Nov. 2014, pp. 109–120

- **Chapter 6:** Chris Dijkshoorn, Cristina-Iulia Bucur, Maarten Brinkerink, Sander Pieterse, and Lora Aroyo. "DigiBird: on the fly collection integration supported by the crowd." In: *Proceedings of the museums and the web conference*. Apr. 2017

I am the main author of all papers listed above: I developed the ideas, collected the data, conducted the analysis and reported upon the findings. Some of the datasets and applications described in these papers have been provided by co-authors. The data analyzed in Chapter 2 stems from a conversion of collection data to Linked Data, created by Lizzy Jongma. I am the main contributor to the codebase of the Accurator annotation tool described in Chapter 4, which incorporates an annotation component developed by Michiel Hildebrand and Jacco van Ossenbruggen. The Accurator annotation tool is used to support and analyze the nichesourcing methodology I developed together with members of the SEALINCMedia project. Jan Wielemaker created the search algorithm used to analyze the impact of enrichments on search result diversity in Chapter 5. I was the lead developer of the DigiBird collection integration framework, discussed in Chapter 6.

2

THE RIJKSMUSEUM COLLECTION AS LINKED DATA

Many cultural heritage institutions provide online access to collections. It is beneficial for institutions to publish their datasets as Linked Data, in order to achieve easy interlinking and integration. In this chapter, we analyze the Linked Data of the Rijksmuseum. We provide collection and vocabulary statistics, as well as challenges encountered during the conversion process. At its time of publication, in March 2016, the Linked Data version of the collection contains over 350,000 objects, including detailed descriptions and references to high-quality images released under a public domain license. Although the number of available object descriptions is rising, we conclude from the analysis that less than half of the Rijksmuseum collection is available as Linked Data. Furthermore, the chosen data model simplifies the source data, thereby omitting information essential for some use cases. And while many object descriptions are contextualized using references to external datasets, many additional links could be added. Therefore, we discuss the impact of data models on published data in Chapter 3 and enrich subsets of the Rijksmuseum collection in two nichesourcing case studies of Chapter 4.

This chapter was published as "The Rijksmuseum Collection as Linked Data" in the Semantic Web Journal (Dijkshoorn et al. [22]) and was co-authored by Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele and Jan Wielemaker.

2.1 INTRODUCTION

Publishing cultural heritage collections as Linked Data improves reusability of the data and allows for easier integration with other data sources [72, 90]. Concepts providing context for collection objects are often shared among multiple cultural heritage institutions, which is an ideal basis for creating connections between collections and allowing reuse of information [43, 77]. The availability of data models tailored towards publishing cultural heritage data helps to make the data available in an interoperable way [26, 27]. These benefits have become apparent to the sector, resulting in an increase of attention and the development of methodologies to help institutions overcome the hurdles involved in publishing data according to the Linked Data principles [42, 72, 90].

The Linked Data version of the Rijksmuseum collection has some unique features. The data is a result of a joint effort between the mu-

seum, the Centrum Wiskunde & Informatica and the Vrije Universiteit Amsterdam and has evolved with input from many research projects [3, 67, 81]. Nowadays, employees of the museum are in control of the publication process, creating and maintaining a conversion layer from the collection management system to Linked Data. The museum's digitization process includes the use of external datasets for adding contextual concepts (e.g. subject or material), creating manually curated links towards structured vocabularies [39]. The data is continuously extended: every day new objects and descriptions are added and both metadata and images are released under open licenses when possible.

This chapter describes the Rijksmuseum Linked Data and provides insights into the lessons learned during its creation. The lessons learned regarding data modeling and contextualization serve as input for subsequent chapters. In the next section, we describe the characteristics of the Rijksmuseum collection and its digitization process. The historical development of the dataset is given in Section 2.3. Section 2.4 concerns an analysis of the state of the dataset as of 2016. This section includes a description of the conversion approach, an analysis of the data model, dataset statistics and an overview of the links from collection objects to external data sources. In Section 2.5 we illustrate uses of the data, before we conclude in Section 2.6 with a discussion.

2.2 THE RIJKSMUSEUM IN A DIGITAL AGE

The Rijksmuseum Amsterdam is one of the most visited museums in the Netherlands, with a mission to provide a representational overview of Dutch art from the Middle Ages onwards. It is well known for its Golden Age paintings, including artworks by Rembrandt and Vermeer. The collection comprises over a million objects, of which only a fraction can be on display at a given time. To open up the remaining collection the museum started digitizing objects and publishing them online.

Digitizing large collections is a time consuming and costly endeavor. To address the backlog of objects to be digitized, the Rijksmuseum started a dedicated digitization project, employing catalogers and professional photographers. The catalogers register objects in the collection management system and add metadata, using structured vocabularies if available [39]. The photographers take high-quality pictures which are released under a public domain license when possible, waiving the rights of the museum.

The digitized collection objects are accessible through the website of the museum. Online visitors can explore the collection using categories or they can search for specific keywords. The presentation of the website focusses on high-quality images of collection objects, en-

couraging users to save, manipulate, and share them [34]. Developers can use an Application Programming Interface (API) to get access to information about the collection objects, sub-collections created by users, and event information[1].

2.3 HISTORY OF THE RIJKSMUSEUM LINKED DATA

The Linked Data version of the Rijksmuseum dataset has a long history, influenced by a number of research projects. A first Resource Description Framework (RDF) version comprising 750 top pieces was created by converting a data dump from an educational database [31]. As a next step, in an effort to integrate Dutch cultural heritage collections, the data model was changed to follow the VRA Core specification[2], with the key advantages of allowing the use of Dublin Core constructs[3] and making a distinction between the physical object and its digital representations. The metadata values of objects were represented in plain text.

In a next version, contextual concepts from in-house thesauri of the Rijksmuseum were aligned with the Getty thesauri[4] and WordNet[5], resulting in a dataset of 27,993 triples [67]. At the time, the Getty vocabularies were only available under license and in the XML format, which resulted in the need for an internally maintained conversion to RDF. In a similar effort, the vocabulary Iconclass was converted and aligned using the Simple Knowledge Organization System (SKOS) to formalize its structure [77]. The experiences gained served as input for the SKOS specification.

The Rijksmuseum dataset was one of the first entries in the Europeana Thought Lab[6], an initiative for showcasing experimental technologies. This entry marks the first conversion of all available Rijksmuseum collection data: 46,000 objects with images were obtained from the collection management system and converted to comply with the VRA data model. The experience of modeling the complete collection and integrating it with collections from other institutions required the ability to model different (potentially conflicting) metadata records from different sources describing the same object. These and other gathered requirements influenced the creation of the Europeana Data Model [27].

The Europeana Data Model today has a set of core and contextual classes that can capture collection information. The data model is designed with reuse of existing classes and properties in mind. It includes elements from the Dublin Core metadata initiative and the

1 http://www.rijksmuseum.nl/en/api
2 http://www.loc.gov/standards/vracore/schemas.html
3 http://dublincore.org/
4 http://www.getty.edu/research/tools/vocabularies/lod/
5 http://www.w3.org/TR/wordnet-rdf/
6 http://labs.europeana.eu/apps/SearchEngineEuropeana

Object Reuse and Exchange definition of the Open Archives Initiative[7]. Cultural heritage organizations can extend the set of classes and properties when needed, reusing elements of other data models or by defining their own. The possibility of making the collection available on the Europeana portal led to the museum taking matters into their own hands. We describe the current conversion of collection data into Linked Data in the next section.

2.4 THE LINKED DATA OF THE RIJKSMUSEUM

In this section we analyze the Rijksmuseum Linked Data, starting with the conversion approach in Section 2.4.1. Sections 2.4.2 and 2.4.3 provide details on the data model and the number of digital objects currently available. In Section 2.4.4 we give an overview of the links from collection objects to external data sources.

2.4.1 *Conversion of collection data into Linked Data*

To create Linked Data, a conversion needs to take place from the data contained in the collection management system into RDF. As of March 2016, the collection management system includes 597,193 registered objects which can be described using 597 available fields. Multiple steps are taken to select and convert a subset of fields and objects, which we will describe in the remainder of this section.

Data from the collection management is harvested daily and loaded into a database which serves the website. Not all of the 597 available metadata fields are included in the output of the collection management system, a subset of 245 fields is specified in a dedicated file[8]. Fields that are no longer used or contain sensitive data such as insurance values are excluded. The selected fields are transformed to form field names which better reflect their content, omit empty values and generate links to other databases maintained by the Rijksmuseum. This conversion is accomplished using an Extensible Stylesheet Language Transformations (XSLT) file[9].

On top of the database runs an API, which is used for outputting RDF. Not all of the 597,193 registered collection objects are included in the output, a subset is selected based on copyright statements and the ownership of the object. This results in a set of 351,814 objects which are under the management of the museum and are free of rights. Whether a collection object is under the management of the museum is loosely defined, it includes objects owned by the museum,

7 `http://www.openarchives.org/ore/`

8 `http://github.com/Rijksmuseum/conversion_adlib` includes file adlibweb.xml which identifies the metadata fields that are included.

9 `http://github.com/Rijksmuseum/conversion_adlib` includes file rijksstudio.xslt which transforms the data.

the state and the city of Amsterdam, but also objects which are on permanent loan. Objects which are on loan for a period shorter than six months are not included.

Selected collection objects are converted into RDF with a second XSLT file[10]. Every relevant metadata field of a collection object is mapped to a property of the Europeana Data Model that most closely resembles the values of the field. Since many of these properties originate from the Dublin Core Metadata Initiative, they often describe the data in more generic terms as the original field, causing a loss of precision. We describe the resulting data model for the Rijksmuseum collection in the next section.

For some fields only textual values are available, others are described using contextual concepts. These concepts are manually added in the collection management system by employees documenting the collection objects. Employees can select concepts from a combination of Rijksmuseum thesauri and external datasets and all concepts have a unique identifier. If for selected fields such an identifier is encountered during the conversion, a reference to the resource is added as well as text in the form of the label of the resource.

The output of the API is used to obtain a complete harvest of the data, which is in turn loaded into a graph database (i.e. a triple store). These harvests are run on a monthly basis by an employee of the museum, who updates the triple store by loading the latest version and who provides links to downloads of older data dumps, which are versioned according to the year and month they were obtained. The file `201603-rma-edm-collection.ttl.gz` is used to obtain statistics in Section 2.4.3.

2.4.2 *Data model and URIs*

The Linked Data version of the Rijksmuseum collection is modeled according to the Europeana Data Model (EDM). EDM reuses elements from existing models such as Dublin Core. The structure of the model is expressed with RDF Schema, using constructs like subclass and sub-property relations. The Web Ontology Language is used to relate EDM elements to other data models.

The data model makes a distinction between a collection object and its digital representation(s). This is achieved with three core classes: *edm:ProvidedCHO* for cultural heritage objects, *edm:WebResource* for web resources and *ore:Aggregation* for aggregations of resources. Figure 2 shows the metadata of a painting by Rembrandt, including its core and contextual classes. We use this example throughout the section.

10 `http://github.com/Rijksmuseum/conversion_oai_formats` includes file europeana_edm.xslt which provides an overview of the mappings.

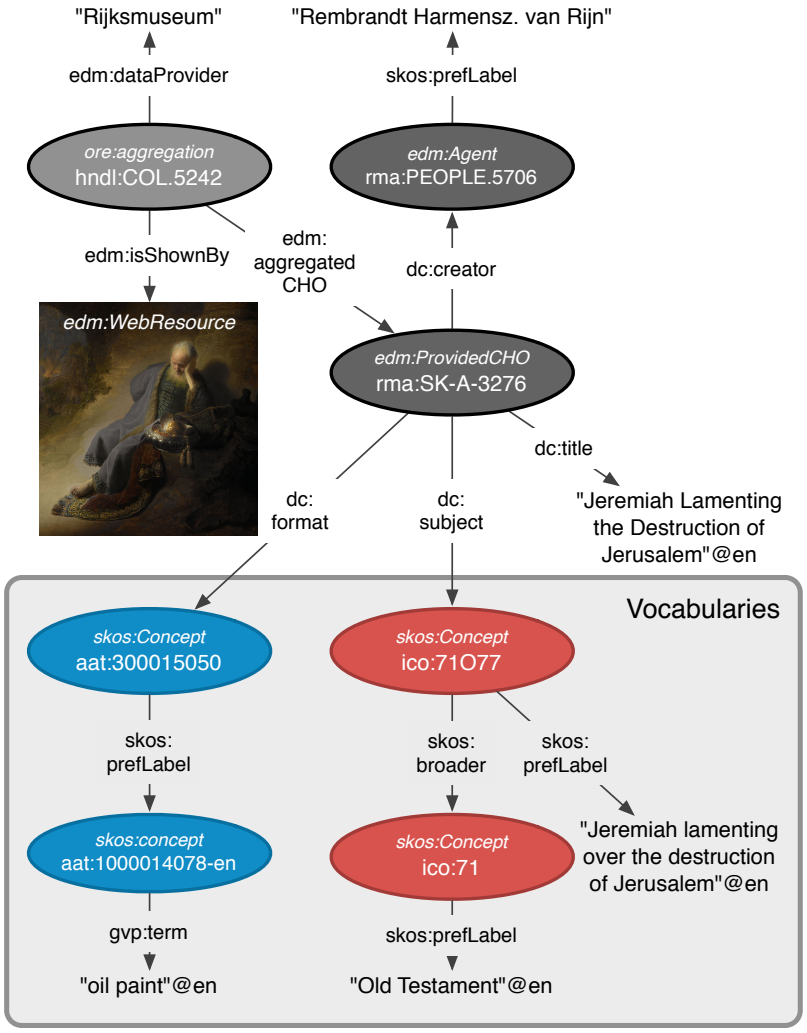

Figure 2: Example of the painting "Jeremiah Lamenting the Destruction of Jerusalem" modeled according to the EDM data model.

An *ore:Aggregation* is used to connect the metadata of a cultural heritage object to web resources. Every collection object in the collection management system gets an aggregation resource with its persistent identifier as URI. Information can be added to the *ore:Aggregation*, Figure 2, for example, shows that the Rijksmuseum served as data provider.

Every *ore:Aggregation* is connected to a resource of class *edm:ProvidedCHO*, representing a description of the physical cultural heritage object. Figure 2 shows four of the properties used to describe objects in the Rijksmuseum dataset: *dc:creator*, *dc:title*, *dc:format* and *dc:subject*. When possible, concepts are used to describe aspects of the artwork, such as the thesaurus term *purl:PEOPLE.5706* for Rembrandt and the concept *aat:300015050* for oil paint. Section 2.4.3 lists the occurrences of predicates used to describe objects in the Rijksmuseum dataset.

When a digital representation is available, the aggregation points to the URL were the image can be obtained. This URL is of type *edm:WebResource* and can, in turn, be described with metadata, adding for example information about its creator. Note that the creator of the image most often differs from the creator of the artwork. The Rijksmuseum dataset currently includes information about the date of creation and the file format of the image.

Not all intricacies of the collection data can be captured using constructs of the Europeana Data Model. While the source data includes detailed information about creator roles and fields like "rejected creator", no such properties exist in Dublin Core. We discuss these data modeling challenges in more depth in Chapter 3. EDM allows for refining and extending the data model, so in the future, the museum can choose to introduce its own more specific constructs or find others to reuse. This could increase the coverage of data in the collection management system included in the Linked Data version.

Persistent identifiers in the form of handles[11] are used for the URIs of the *ore:Aggregation*. Since an aggregation connects metadata of the object and its digital representation, the persistent identifier is not related to the object number. The URI of the cultural heritage object descriptions is based on the purl scheme[12] and consist of five elements: purl prefix, dataset type, country code, organization and object number. This results in the following URI for the *edm:ProvidedCHO* resource of the Rembrandt in Figure 2: `http://purl.org/collections/nl/rma/SK-A-3276`. When values refer to one of the thesaurus databases of the museum, a URI is generated based on the internal reference used, linking the collection object with the thesaurus.

11 `http://www.handle.net/`
12 `http://purl.org/`

Table 1: Overview of the predicates that describe collection objects. Resources and literals refer to the distinct objects used to describe collection objects, the language codes indicate the availability of literals in given languages.

predicate	artworks	resources	vocabs	literals	en	nl
dc:contributor	89,796	0	-	9,146	0	9,146
dc:coverage	138,141	0	-	212	106	106
dct:created	340,865	0	-	46,255	19,837	19,837
dc:creator	349,787	27,904	rma	38,851	97	38,754
dc:description	217,202	0	-	176,786	2,487	174,299
dct:extent	288,318	0	-	55,241	15,229	40,012
dc:format	322,152	593	aat, rma	924	322	602
dct:hasPart	487	0	-	13,646	0	0
dc:identifier	351,814	0	-	703,429	0	0
dct:isPartOf	59,197	0	-	2,754	0	0
dct:isRefBy	98,815	0	-	80,770	0	0
dc:language	351,814	0	-	1	0	0
dct:prov	13,239	0	-	1,155	0	1,155
dc:publisher	351,814	0	-	1	0	0
dct:spatial	214,752	3,339	rma	3,361	0	0
dc:subject	221,868	44,840	ic, rma	32,452	0	32,452
dc:title	351,789	0	-	297,271	6,784	290,487
dc:type	351,749	3,541	aat, rma	3,700	155	3,545
edm:type	351,814	0	-	1	0	0

2.4.3 *Rijksmuseum dataset statistics*

As of March 2016, the Linked Data version of the Rijksmuseum collection[13] comprises 22,846,996 triples, describing 351,814 objects, of which 207,441 have a graphical depiction. Ten sub-collections are maintained, including sculptures (29,782 objects), historical items (19,936 objects), paintings (3,949 objects) and Asian art (3,722 objects). The collection of works on paper has 280,047 objects and is by far the largest sub-collection, including prints, drawings and photos.

Table 1 lists the predicates used to describe collection objects. A title is provided for almost all objects, of which the majority is unique. Although most of the titles are in Dutch, some of them are also available in English. Over half of the objects have the predicate *dc:description*, which includes textual information about the subject matter and art-historical background of the object. For example, the description of Figure 2 includes the following text: "Downcast, the biblical prophet Jeremiah leans his tired head on his hand" and "Rembrandt used powerful contrasts of light and shadow to heighten the drama of the scene".

13 `http://datahub.io/dataset/rijksmuseum`

There are over 30,000 unique creators, which are mostly described using resources from the person database of the museum. Half of the *dc:creator* literals are based on labels from resources in the person database, while the other half is used for adding nuances to the creator field which are difficult to capture in a resource of type *edm:Agent*. This includes textual descriptions such as "Anonymous", "possibly Rembrandt" and "follower of Rembrandt". The predicate *dc:contributor* refers to names of additional persons involved in the creation process.

The *dc:subject* predicate provides information about the subject matter, where resources from both the Iconclass vocabulary as well as the Rijksmuseum thesaurus are used. Subjects are also described using Dutch literals, since not all subject matter falls within the scope of the available vocabularies. The predicate *dcterms:spatial* refers to places, for which both terms from the thesaurus as well as language agnostic literals are used.

The museum describes temporal aspects using the predicates *dc:coverage* for periods and *dcterms:created* for creation dates. Dutch as well as English literals are used for both predicates, where creation dates are expressed using a year or estimated years (e.g. "1630" or "c.1600-c.1625") and periods are expressed using textual descriptions (e.g. "second quarter 17th century" or "18th century").

The predicate *dc:type* uses a mixture of concepts from the Art & Architecture thesaurus, the museum's thesaurus and literals to identify the type of object (e.g. print or painting). The same applies to *dc:format*, which is used to specify materials such as the resource *aat:300015050*, which stands for "oil paint". In the next section, we describe in more detail how many of such connections are made to external datasets. Physical dimensions of the object are recorded using *dcterms:extent*, specifying the height and width of objects in centimeters. The painting in Figure 2 has for example "height 58 cm" and "width 46 cm".

Objects can be connected to each other with the *dcterms:hasPart* and *dcterms:isPartOf* predicates. These relations are, for example, used to relate photographs to their album. Related sources (often books) are linked to the object using *dcterms:isReferencedBy*. These three predicates currently refer to literals representing identifiers, which in a later stage can be converted to resources matching the objects indicated by the identifiers. Every object has two identifiers, one for internal use (e.g. "SK-A-3276") and one persistent identifier in the form of a handle (e.g. "hndl:RM0001.COLLECT.5242").

The predicate *dc:provenance* encodes the provenance in a literal enumerating the present and past owners of the object. Most of the intellectual property rights are part of the public domain, while sometimes specific persons are specified who own the copyrights. Europeana requires some values to be present and limited to a set range.

Table 2: Types in the Rijksmuseum thesaurus with more than 500 values.

type	distinct resources	nl labels	en labels
person	38,939	27,904	27,904
place	17,174	17,174	152
object name	6,074	6,074	298
keyword	5,021	5,021	166
event	1,982	1,982	43
technique	1,401	1,401	73
occupation	1,044	1,044	15
material	882	882	428
location RMA	808	808	5

The publisher is always the "Rijksmuseum", while *dc:language* is the language code of the country of the institution, in this case, "nl". The *edm:type* is "IMAGE".

As outlined in Section 2.4.1, some of the literals are based on the labels of resources. Although adding both the literal and resource introduces redundant information, this can support applications that do not handle the added complexity of resources well. The literals from *dc:type*, *dc:format* and *dcterms:spatial* all directly originate from the museum's thesaurus. 77 percent of the literals of the *dc:subject* field come from either resources contained in the thesaurus or the person database. The remainder of the subject literal values mainly describe specific dates and periods such as "1701 - 1703". We describe the resources contained in the thesauri and links to external datasets in the next section.

2.4.4 *Contextual concepts and links to external datasets*

Institutions often maintain their own vocabularies containing their perspective on contextual concepts. When the contextual concepts of collection objects are replaced with concepts from standardized vocabularies such as the Getty vocabularies, these nuances in perspectives are in danger of disappearing. So while collection objects and contextual concepts in the thesauri of the Rijksmuseum are linked to an increasing number of available datasets maintained by other institutions, the Rijksmuseum chooses to also maintain and use its own. This allows the museum to preserve its own perspective and in a later stage, vocabulary alignment tools can be used to match the concepts with similar concepts in external datasets [73].

Five contextual classes are defined in the Europeana Data Model for relating collection objects to contextual information: *edm:Agent*, *edm:Place*, *edm:TimeSpan*, and *skos:Concept*. These classes correspond to the types of thesaurus records in the databases of the Rijksmuseum: the person database maps to the agent class and the general the-

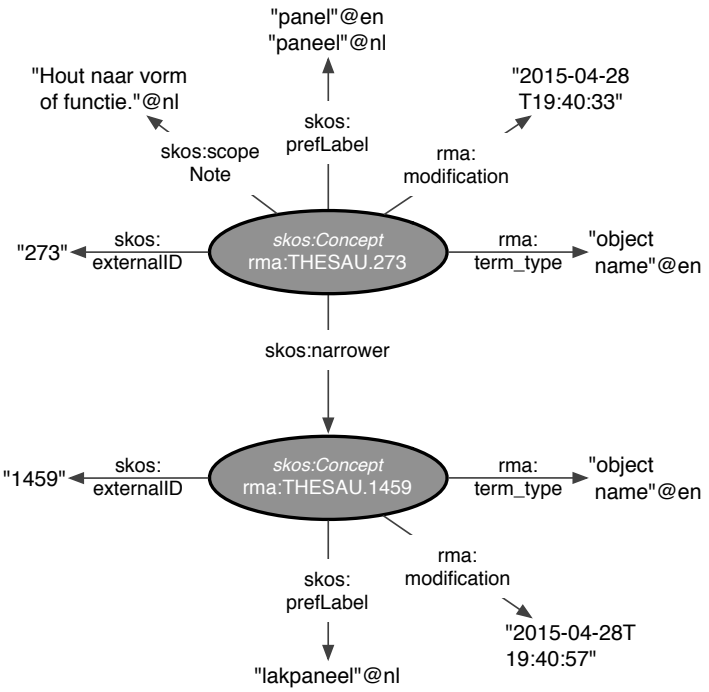

Figure 3: Diagram of the thesaurus term representing "panel".

saurus database contains information about places, historical events, and other concepts. However, the types of concepts in the museum's thesaurus are divided into finer grained types. An overview of the type of concepts is presented in Table 2, along with the number of available resources and labels.

The thesaurus forms a hierarchy of concepts using relations such as broader and narrower, which are represented using SKOS. Figure 3 shows two concepts, where the type of concepts is indicated using the *rma:term_type* predicate. All of the 33,800 concepts in the thesaurus have a Dutch label, only 1,539 have an English label. For 3,254 terms a *skos:scopeNote* is available, describing the appropriate use of the concept. Every concept has its own unique *skos:externalID* and the last modification date of the concepts is recorded using *rma:modification*.

The person database contains resources of type *edm:Agent*, Figure 4 shows "Rembrandt" as an example. The names of persons are indicated using *skos:prefLabel* and every person has a name, either represented as a Dutch, English, or language independent literal. Additional information about persons is added using the Resource Description and Access (RDA)[14] vocabulary. A *rdv:professionOrOccupation* is provided when known, Rembrandt has, for example, multiple listed professions. Besides being a painter he also made prints. When available, information is added about the birth and death of the person. In the remainder of this section, we describe how external datasets are

14 http://rdvocab.info

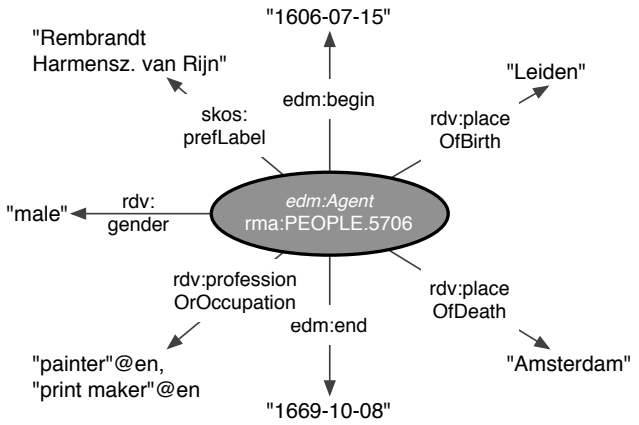

Figure 4: Diagram of the agent resource representing the person "Rembrandt".

used to extend the thesauri and annotate the collection data of the Rijksmuseum.

The **Art & Architecture Thesaurus**[15] (AAT) consists of concepts about arts from antiquity to the present. Concepts include art styles, materials and agents. It is maintained by the Getty Foundation, which released a Linked Data version in February 2014 with 38,619 concepts. The focus of the thesaurus lies on generic concepts: instead of for example describing individual artists, it includes the concept "printmakers". New concepts originate from cataloging and documentation projects and labels of concepts are available in multiple languages.

The Rijksmuseum uses the Art & Architecture Thesaurus for the *dc:type* and *dc:format* metadata fields. A small subset of the available concepts is used: 305 distinct formats and 124 distinct types. As can be seen in the type frequency distribution in Figure 5, a small number of concepts is often used. This is also the case for the format field. For example, the top three types are prints (183,916), stereoscopic photographs (3,480) and plates (1,617). The museum refrains from assigning art styles to objects, since it is often debatable to which art style an object belongs.

The **Iconclass vocabulary**[16] contains 39,578 concepts, providing "a systematic overview of subjects, themes and motifs in Western art". An official Linked Data version was released in 2012. Concepts are identified with codes and SKOS relations are used to create a hierarchy between them. Labels of concepts are available in English, German, French, Finnish and Italian. An example of a code used in Iconclass is 7, which refers to the "Bible" and is connected to the concept 71O7, "the book of Jeremiah", using *skos:narrower* predicates. Context-dependent modifiers can be added to the codes: for 71C131(+3), the

15 http://www.getty.edu/research/tools/vocabularies/aat/
16 http://www.iconclass.nl/

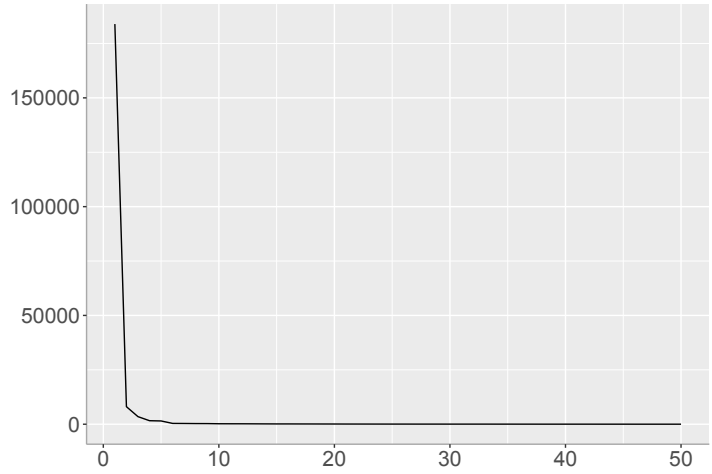

Figure 5: Frequency distribution of the top 50 type concepts of AAT the collection objects are linked to.

code 71C131 indicates "the sacrifice of Isaac", while the modifier (+3) indicates that one or more angels are depicted on the object.

The museum uses the Iconclass vocabulary to describe subject matter. Iconclass codes are added by catalogers during the registration process described in Section 2.2. Out of the 39,578 concepts in the vocabulary, 10,434 are used to add information to an object. Of the 351,814 collection objects, 172,059 have one or more Iconclass annotations. As Figure 6 shows, many of the concepts are often used, while on average a code is used 27 times.

The **Short-Title catalogue Netherlands** (STCN) is "the retrospective national bibliography of the Netherlands in the period 1540-1800"[17], maintained by the National Library of the Netherlands. A Linked Data version is available, containing records of 196,396 publications. This dataset contains many books that are the source of objects in the print collection of the Rijksmuseum and linking the two collections provides valuable contextual information.

The catalogers of the Rijksmuseum add references to the National Library by adding textual descriptions of the books in a notes field. To create links, these descriptions are scanned for objects from the STCN that match the title, publication date and publisher. This automated matching process resulted in 3598 links from the Rijksmuseum collection to 501 publications in the STCN catalogue. The links are encoded as *dc:hasPart* relations from the STCN vocabulary to the Rijksmuseum collection. Catalogers estimate that roughly 14,000 works could have been linked to STCN. A more rigorous way of referring to STCN titles is in the making to support this.

17 `http://www.kb.nl/expertise/voor-bibliotheken/short-title-catalogue-netherlands` (accessed on 04-07-2014)

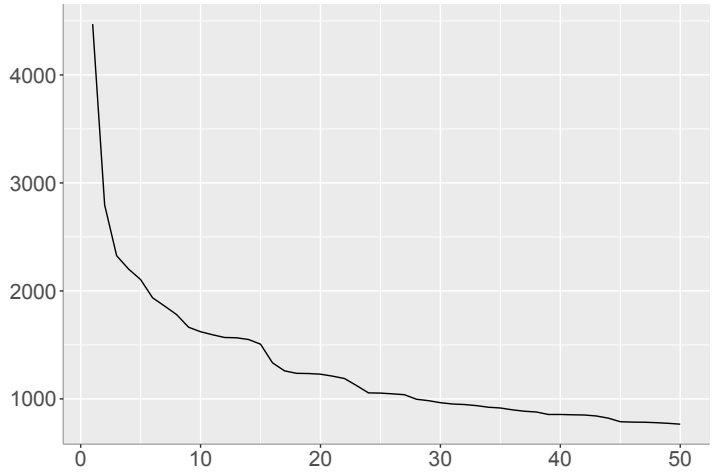

Figure 6: Frequency distribution of the top 50 subject concepts of Iconclass that the collection objects are linked to.

2.5 DATA USAGE

Uses of the Linked Data of the Rijksmuseum include search, recommendation, collection integration and browsing. In this section, we give an overview of how the museum data has been used in various research projects and provide statistics about the Rijksmuseum API. Most projects that contributed to the process of data development had demonstrators illustrating the power of Linked Data.

The **MultimediaN E-Culture** project showcased a semantic search system, which won the 1st price in the 2006 International Semantic Web Conference Challenge [67]. It clustered search results based on the graph path leading from a matching literal to objects. The dataset was extended from 750 artworks to the entire Rijksmuseum collection in a search prototype of the **Europeana Thought Lab**[6], showing advanced search functionality to be included in the portal at a later stage. This same search algorithm is used to investigate the impact of enrichments on search result diversity in Chapter 2 of this thesis.

Other ways of accessing data were introduced in subsequent years. The **CHIP** demonstrator recommended artworks based on graph patterns [81]. The **STITCH** project took a different approach with facets based on Iconclass concepts, allowing users to browse the collection based on different topics [77]. The **Agora** demo provided access to the collection with an emphasis on the events related to objects [3]. As discussed in Chapter 4, the **Accurator** nichesourcing tool uses graph patterns to recommend people artworks to which they can contribute information, gathering more accurate subject matter descriptions [18].

The Rijksmuseum maintains an API[1] for application developers, optionally returning data formatted according to the Europeana Data Model. 587 people have registered for access to the API as of August

2015 and many different applications have been build on top of it[18]. The API is used by Europeana to harvest collection data, making all the structured data of the Rijksmuseum available through the Europeana portal. Europeana logs the page views of this portal and during a period of 20 weeks (starting from the 1st of May 2015), Rijksmuseum collection objects got 42,156 page views of which 34,206 were unique.

2.6 DISCUSSION

For a long time, Linked Data has been a promise for data publication and integration in the cultural heritage sector. Despite widespread interest and apparent advantages, only a limited number of institutions have managed to make their collection available as Linked Data. After a period of development influenced by many research projects, the Rijksmuseum is one of them. Furthermore, the museum is in control of the entire publication process of its own collection as Linked Data.

The majority of the Rijksmuseum collection objects are part of the public domain since their intellectual property rights have expired. Although general understanding is that digitized representations of public domain works should again be released under the same license terms, many institutions are hesitant to do so, in fear of losing a possible revenue stream. Nonetheless, the dataset presented in this chapter includes 207,441 references to images. The Rijksmuseum did release their high-quality images in the public domain in 2013, arguing that the increase in attention and exposure would result in a higher number of visitors [60]. In turn, it allowed the museum to gain more control over the digital representations that had appeared online, replacing many inferior versions by its high-quality images.

The quality and correctness of metadata are of paramount importance to museums [72]. The Rijksmuseum has an extensive quality control process in place to ensure the correctness of metadata. By adding a direct conversion layer to the collection management system it ensures that the same level of quality is translated to the Linked Data version. However, as discussed in Section 2.4.2, the chosen data model impacts the information that can be captured. This aspect is clearly visible for creation events, the roles an agent played now has to be encoded in a literal. In the next chapter of this thesis, we further explore the impact of different modeling approaches on the information that can be conveyed.

Data aggregators such as Europeana enticed many institutions to provide digital versions of their collection, often relying on external expertise for the conversion process. This led to an increase in

18 `http://www.opencultuurdata.nl/category/apps/` provides an overview of applications that use cultural heritage data, including applications that are built on top of the Rijksmuseum API.

available collections, although providing access to data through aggregators has the major drawback that it creates a gap between the institution and its data [90]. This gap can lead to misinterpretation of data and the obfuscation of the source of the data. We believe it is therefore still desirable that institutions publish their own data if the required expertise is available. They thereby remain in control of choosing the most suitable data model, URI naming schemes, links to other datasets, and update processes.

The data of the Rijksmuseum is subject to constant change: newly digitized objects are added on a daily basis and employees extend and refine information regularly. This should not come as a surprise when you realize that the museum's collection consists of over a million objects, of which just 593,193 objects are described in the collection management system. The number of included objects in the dataset described in this chapter is even lower since it only includes objects which are under the management of the museum and that are free of rights. The museum could in the future decide to extend this number by also releasing metadata of objects still under copyright.

The Linked Data version of the Rijksmuseum collection places the objects in a broader context, by relating the object metadata to external datasets. The museum uses the Getty thesauri to describe aspects such as materials and types of objects. Room for improvement remains: persons and places are not yet related to external datasets and not all objects have associated types and materials. Iconclass is used extensively to describe subject matter, although 129,946 of the published object description lack any form of subject description. In Chapter 4, we investigate how the accuracy and coverage of object descriptions can be extended using nichesourcing, thereby adding new perspectives to the data and increasing the possibilities of meaningful reuse.

3

MODELING CULTURAL HERITAGE DATA FOR ONLINE PUBLICATION

In this chapter, we investigate how cultural heritage data can be contextualized. Ontologies are used to structure data, while defining the contextual information that can be conveyed about objects. We formulate requirements for cultural heritage ontologies, based upon our experience of publishing data of the Rijksmuseum and related work. The requirements regard specialization, object- and event-centric approaches, temporality, representations, views and subject matter. For each requirement, common modeling approaches are investigated, by discussing two models regularly used in the museum sector: the CIDOC Conceptual Reference Model and the Europeana Data Model. The outlined approaches and requirements provide insights into data modeling practices reaching beyond the cultural heritage sector.

This chapter was published as "Modeling Cultural Heritage Data for Online Publication" in the Applied Ontology Journal (Dijkshoorn et al. [21]) and was co-authored by Lora Aroyo, Jacco van Ossenbruggen and Guus Schreiber.

3.1 INTRODUCTION

Cultural heritage institutions possess a wealth of information and there is a growing understanding that it would be beneficial to share this online. It is, however, a non-trivial step for institutions to move from traditional information dissemination methods, such as exhibition catalogs and research papers, to the publishing of data [50]. To do so, cultural heritage institutions need to consider aspects that were previously not part of their core activities (e.g. using standardized data models and aligning data with other institutions). To aid this transition, this chapter will: 1) outline data models in the cultural heritage domain; 2) demonstrate using two example artifacts, that different modeling approaches impact the information that can be published; and 3) combine insights from modeling challenges to form requirements, in order to aid institutions in choosing an appropriate data model.

Most cultural heritage data is contained in collection management systems and catalogs that often function as data silos; systems that can not be accessed by the outside world [43]. Over the years, efforts to export this data have had many manifestations, resulting in different data formats and data models. This makes the reuse and integration of cultural heritage data a cumbersome task. Adoption of

Linked Data practices improves the interoperability of cultural heritage data. Linked Data principles advocate systematic referring to resources and syntactically standardized publishing of data [8]. Ontologies make the semantics of published data explicit, by providing a shared conceptualization [71]. The reuse of ontologies is encouraged and there are specialized ontologies available for many different domains [48]. The cultural heritage domain is no exception, with models specifically tailored to libraries, archives and museums [25].

A cultural heritage institution is confronted with many important choices when it decides to publish Linked Data and reuse an ontology [56]. One of the first decisions to make is whether to invest in infrastructure to publish Linked Data or provide data to an aggregator of cultural heritage information. The former leaves the choice of which ontology to use up to the institution, while an aggregator might require data to be provided in a specific structure [13]. The decision of which ontology to use has implications for the source data that can be included, as well as the shape of the resulting Linked Data. In this chapter, we discuss the impact of such decisions, with respect to six modeling challenges. The challenges originate from related work, as well as our experience of publishing data about objects in the collection of the Rijksmuseum Amsterdam (Chapter 2).

We illustrate different modeling approaches using two cultural heritage ontologies. The CIDOC Conceptual Reference Model (CIDOC-CRM) is specifically developed for the museum sector and is intended to be used to create interoperable data. The Europeana Data Model (EDM) is an ontology that enables cultural heritage institutions to structure collection data so that it can be used by the data aggregator Europeana. This model is designed to retain other more specific data models that are used by libraries, archives and museums. We chose these ontologies for illustrating modeling approaches since they are commonly considered by institutions for publishing data, but take a different approach in doing so. We draw requirements for cultural heritage ontologies from the challenges.

The remainder of this chapter is structured as follows: in Section 3.2 we discuss related work on cultural heritage ontologies and datasets, to identify modeling challenges encountered while publishing cultural heritage data. We describe two exemplary artifacts of the Rijksmuseum in Section 3.3: a wedding portrait and a cased pair of pistols. These objects are suitable for illustrating modeling challenges, because of their distinct types, related events and subject matter. The discussed modeling challenges regard specialization, object- and event-centric approaches, temporality, representations, views and subject matter. The challenges and modeling approaches addressing them, are discussed in Section 3.4. We discuss requirements distilled from these challenges in Section 3.5 and end with a conclusion and future work section.

3.2 RELATED WORK

In this chapter, we focus on how ontologies can be used to structure and represent information about artifacts in cultural heritage collections. In the context of computer science, Studer, Benjamins, and Fensel [71] defined an ontology as a "formal, explicit specification of a shared conceptualization". A conceptualization is an abstract view of the world we want to represent. Making the conceptualization explicit entails deciding on a language to use and constraining the interpretations of such language. "Formal" refers to the specification being machine readable [38]. In Section 3.4, we use the Europeana Data Model and the CIDOC Conceptual Reference Model to illustrate different approaches to modeling cultural heritage data.

The CIDOC Conceptual Reference Model (CIDOC-CRM) is an event-centric reference ontology for the cultural heritage sector, maintained by a special interest group of the ICOM international committee for documentation [25]. Constructs of the CIDOC-CRM are based on empirical studies of collection management systems [15, p. i]. The ontology aims to be a "discipline neutral", common semantic reference point, improving the semantic and structural interoperability of cultural heritage data. CIDOC-CRM has been accepted as an ISO standard for the interchange of cultural heritage information in 2006.

The Europeana Data Model (EDM) is developed to represent and structure cultural heritage data so that it can be delivered to the data aggregator Europeana [27]. EDM is used internally by Europeana to aggregate, process, enrich and disseminate data. The model reuses constructs from other data models, such as the Dublin Core metadata initiative[1] and the Open Archives Initiative Object Reuse and Exchange standard[2] [46]. EDM is a top-level ontology to which institutions can map their more specific data models (CIDOC-CRM can, for example, be embedded in EDM). This approach makes it useful beyond its original purpose of data delivery: the model is nowadays used by other aggregators, as well as institutions publishing their own data [76].

Both ontologies are used by museums that publish a Linked Data version of their collection. For instance, the Amsterdam museum ontology and VVV ontology specialize the EDM top-level ontology to structure collection data [28, 90]. In addition, the Rijksmuseum dataset is published using a combination of the Dublin Core model and EDM, as discussed in Chapter 2. A collaboration of 14 American art museums mapped collection data to CIDOC-CRM [50]. This ontology is also used to publish collection data of the Yale Center of British Art[3], British Museum[4] and Russian Museum [54].

1 http://dublincore.org
2 http://openarchives.org/ore
3 http://collection.britishart.yale.edu
4 http://collection.britishmuseum.org

Figure 7: Portrait of Marten Soolman, created by Rembrandt in 1634.

There are many more cultural heritage collections available as Linked Data, through so-called "aggregators". Aggregators are organizations that host multiple collections, creating an integrated point of access to artifacts of different institutions. The MultimediaN E-Culture project provides integrated access to three cultural heritage collections [67]. MuseumFinland is a collaboration of Finnish museums, while the LODAC Museum includes many Japanese museums [43, 53]. Europeana is an aggregator of European cultural heritage data, which connected data of over 3,000 institutions in 2017 [29].

3.3 TWO EXAMPLES: A PORTRAIT AND A PAIR OF PISTOLS

In this section, we introduce two artifacts from the Rijksmuseum Amsterdam. The first artifact is part of a pair of wedding portraits and the second is a set of pistols. These artifacts together are a good combination for illustrating modeling challenges, due to their distinct types, a range of related events, a number of different representations and an assortment of subject matter. Based on related work and our own experience modeling the Rijksmuseum data, we illustrate typical mod-

Figure 8: Rembrandt exhibition in 1956, including the portrait of Marten Soolman.

eling challenges of the cultural heritage domain with these artifacts in the subsequent section.

PORTRAIT OF MARTEN SOOLMANS The first artifact that we use as a running example throughout this chapter, is the portrait of Marten Soolmans, as depicted in Figure 7. In 1634, Rembrandt van Rijn painted a pair of portraits in honor of the wedding of Marten Soolmans and Oopjen Coppit. The paintings show the young married couple in exuberant detail, dressed in black and adorned with many lace details. While part of private collections for ages, in an exceptional construction, the Dutch and French governments managed to acquire both paintings. The portraits will always be exhibited together and their location will alternate every five years between the Rijksmuseum and the Louvre. The Rijksmuseum maintains an archive, which includes files that document exhibitions. Figure 8 shows a picture of the two marital portraits exhibited at the Rijksmuseum in 1956, as part of an exhibition in honor of the 350th anniversary of the birth of Rembrandt. Two institutions with different views on the same artifact, related events and the availability of digital representations, makes modeling information about the portrait of Marten Soolmans an illustrative example of the challenges that can be encountered during the publication of cultural heritage data.

CASED PAIR OF PISTOLS The second example is a cased pair of pistols, shown in Figure 9. These flintlock pistols were manufactured in the workshop of Jean Le Page, around the year 1808. The pistols lend their historical significance by reputedly being owned by Napoleon I Bonaparte, emperor of France. After the battle of Water-

Figure 9: A cased pair of pistols, reputedly owned by Napoleon.

loo, the cassette containing the pistols was found in the traveling carriage of Napoleon and there are letters supporting the assessment that Napoleon once owned the cassette. Besides the pistols, the cassette also contains accessories, such as a powder horn, bullet mold, rammer and hammer. The pistols are made from multiple materials, such as walnut, steel and gold. The weapons are adorned with engravings both written, as well as figurative. For example, an eagle is depicted on the side of a pistol, while its barrel is engraved in gold, with the text "Arger de l'Empereur". The pistols are well suited for illustrating the challenges of modeling cultural heritage data, because of their components that have been created at different moments in time, the detailed provenance information and the depicted subject matter. But the first challenge we discuss in Section 3.4 is how to differentiate between pistols and portraits, by specializing data models.

3.4 MODELING APPROACHES IN THE CULTURAL HERITAGE DOMAIN

In this section, we discuss six modeling challenges, as listed in Table 3. For each challenge, we discuss the general issues in more depth and show current modeling approaches of the Europeana Data Model and

Table 3: Overview of modeling challenges and approaches.

Challenge 1	How to specialize an interoperable data model?
Approach 1A	Specialization by extending a top-level class hierarchy.
Approach 1B	Typing using terminologies.
Approach 1C	Specialization by extending a property hierarchy.
Challenge 2	How to choose between an event- and object-centric approach?
Approach 2A	Event-centric approach.
Approach 2B	Object-centric approach.
Challenge 3	How to capture changes over time?
Approach 3A	Textual descriptions of changes over time.
Approach 3B	Embedding temporal information in properties.
Approach 3C	Recording temporal information using events.
Challenge 4	How to describe representations of an artifact?
Approach 4A	Aggregations that connect artifact and representation.
Approach 4B	Similarity of artifact and representation.
Challenge 5	How to model multiple metadata sources with alternative views?
Approach 5A	Proxies.
Challenge 6	How to contextualize artifacts based on subject matter?
Approach 6A	Unbound subject matter.
Approach 6B	Subject matter limited by range.

the CIDOC Conceptual Reference Model. From both ontologies, we analyze encodings that are compatible with Linked Data, to which we will refer to as data models. We illustrate the discussion using the two examples introduced in the previous section.

3.4.1 *Challenge 1: how to specialize an interoperable data model*

Libraries, museums and archives hold many different types of artifacts. The Rijksmuseum alone has ten different sub-collections, ranging from paintings to furniture. To achieve the desired level of interoperability across these sub-collections, some level of abstraction is needed to support descriptions on a more generic level. An overly generic data model, however, might "trivialize" descriptions of these distinct artifacts. This happens when important, but collection-specific information is systematically left out because it does not "fit" the general data model used. This has the additional risk that curators and domain specialists stop to support data publishing if they are under the impression that this implies committing to generic models that do not fit their domain sufficiently. At the same time, there are use cases which require generic descriptions of artifacts, for example, to achieve interoperability with collections with slightly different characteristics. This interoperability is important from a managerial perspective, allowing the participation in vertical as well as horizontal integration projects. In relational or XML-oriented data models, users find it often hard to specialize data models without losing their interoperable generic structures. But even when Semantic Web tech-

nologies are used, there are still choices to be made on *how* to provide such collection-specific aspects.

3.4.1.1 *Approach 1A: specialization by extending a top-level class hierarchy*

One approach is to develop a commonly agreed upon, top-level class hierarchy, that provides the required level of abstraction and interoperability but allows more specific descriptions by refining the generic classes given. CIDOC-CRM defines 82 of such top-level classes, whereas EDM describes 18 classes. The documentation of EDM recommends to use the most specific construct available, thereby contributing to the precision of descriptions [45, p. 11]. However, the most specific EDM class that can be assigned to the pistols and painting is the fairly general *provided cultural heritage object*. CIDOC-CRM is a reference ontology with a similar approach and the most specific class that can be assigned to a cultural heritage object is *physical man-made thing*. Neither of these two general classes allows us to differentiate between a painting and a pistol.

This means that for many purposes, institutions may wish to add more specific classes, either by using a shared profile or by using an institute-specific set of extensions. An institution could, for example, introduce the class *weapon* and relate it to the CIDOC-CRM class *physical man-made thing*. Once the class *weapon* is assigned to one of the pistols, it can still be deduced that it is a *physical man-made thing*. At the moment a class *painting* is added as well, it is possible to differentiate between the two types of artifacts.

3.4.1.2 *Approach 1B: typing using terminologies*

Additional typing of instances with terms from hierarchically structured vocabularies is another common approach to achieve specialization without sacrificing generality. These structured vocabularies can take different forms, such as thesauri, classification schemes and gazetteers [42, Chapter 4]. An example is the Art and Architecture Thesaurus[5] (AAT), which is used to relate artifacts to materials and techniques. These structured organizations of concepts serve as shared vocabularies for data publishers and can improve artifact retrieval tasks [5, 86].

Both CIDOC-CRM and EDM include the property *has type*. Using this property we can state that one of the pistols is of type "flintlock pistol". Such terms can often be reused, for example, from structured vocabularies such as the AAT. Many of these vocabularies are structured using the Simple Knowledge Organization System (SKOS)[6]. The terms in a SKOS vocabulary are connected using broader and

5 http://www.getty.edu/research/tools/vocabularies/aat/

6 http://www.w3.org/TR/skos-reference/

narrower relations, forming a hierarchy. Using this hierarchy it is possible to deduce that a flintlock pistol is a more specific term than a weapon. This typing of instances using terms does not, however, impact the more formal RDF or OWL instance/class semantics, nor does it limit the connections that can be made between different instances.

3.4.1.3 *Approach 1C: specialization by extending a property hierarchy*

Properties are used to relate instances to other instances or literal values. CIDOC-CRM includes a total of 262 property definitions, where the EDM definition defines 35 properties and refers to 40 properties of other data models. Properties can also form hierarchies, which in turn can be extended. When extending the property hierarchy, it is important that the meaning of a sub-property is subsumed by that of properties higher up in the hierarchy. An institution could, for example, introduce the property *was painted for*, to relate the portrait of Marten to the wedding. This property should be a sub-property of *was made for* and not of *was used for*. While both properties relate things to activities, *was used for* is a sub-property of *was present at*, something which is not necessarily true for the wedding portrait.

Domains and ranges can be added to properties, thereby indicating which types of instances a property can relate. In Figures 10 and 11, the texts in ovals serve as indicators of these domains and ranges. The EDM property *was present at*, for example, relates *information resources*, *things* and *agents* to *events*. Take, for example, the following statement: "the pistol was present at the battle of Waterloo". The range of the property *was present at* indicates that the instance *the battle of Waterloo* should be of class *event*. Inconsistencies can occur if the wrong properties are extended or aligned and reasoning is used to deduct additional information, something which we will discuss in more depth below.

3.4.1.4 *Ontological commitment and alignment inconsistencies*

A minimal ontological commitment assures maximum reusability [71]. 63 Properties described in the EDM specification do not have a full domain and range specification, where every one of the 262 CIDOC-CRM properties has a domain and range specified. The omission of domain and range specifications makes the ontological commitment of EDM lower than CIDOC-CRM. While this allows EDM properties to connect multiple classes of instances, it refrains reasoners from automatically deducing the types of instances occurring as a domain or range of an EDM property. The ontological commitment of EDM is, however, impacted by alignments with constructs of CIDOC-CRM. As we will see from the example below, this can lead to undesired inconsistencies in the data when reasoning is used. Institutions that

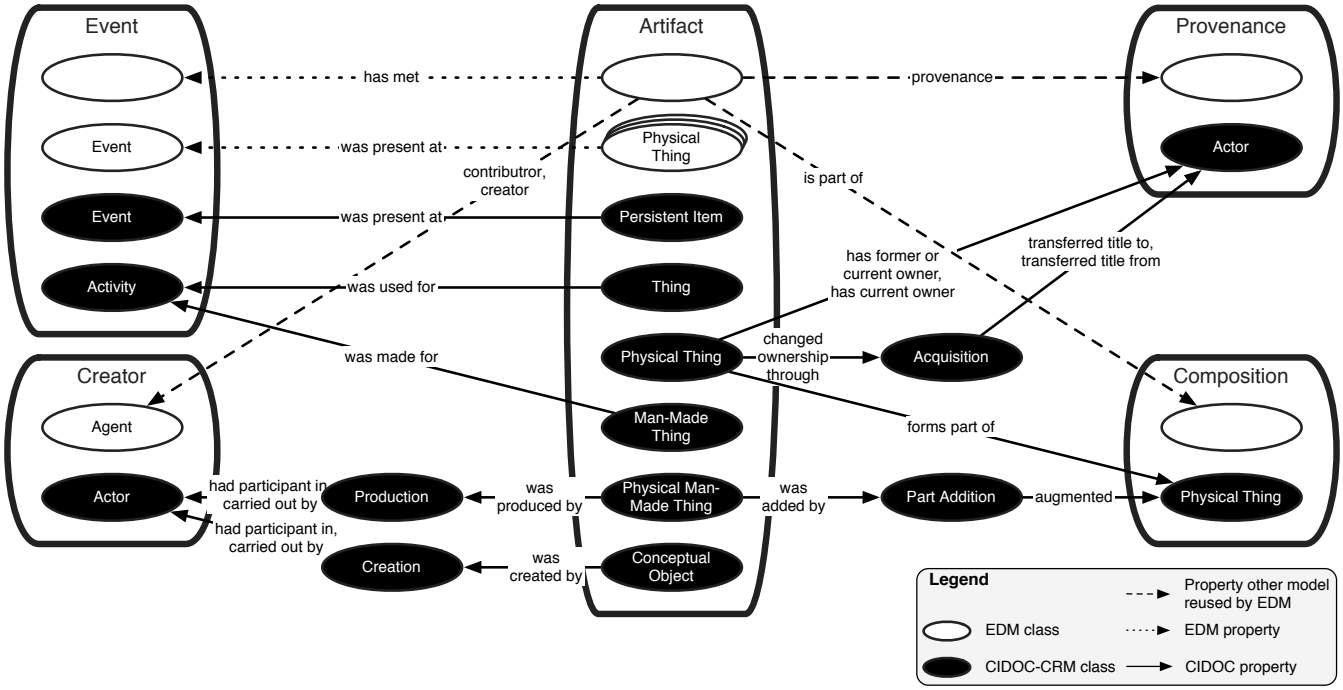

Figure 10: Representations of four elements of an artifacts life cycle: event, creator, provenance and composition.

relate properties to EDM or CIDOC-CRM should consider the ramifications of their alignments carefully.

When new constructs are related to existing data models, care should be taken not to create inconsistencies. The specification of EDM aligns six classes and seven properties with constructs of CIDOC-CRM. An example of problems caused by an inconsistent alignment is the EDM property *is successor of*. This property has no defined domain or range, which allows someone to state that the book the Two Towers is the successor of the Fellowship of the Ring, but also that Queen Elizabeth II is the successor of King George VI. The EDM property *is successor of* is a sub-property of *is similar to*. EDM aligns this property with the CIDOC-CRM property *shows features of*. The latter has a domain and range of thing. Through reasoning, we can now deduct that the domain and range of the property *is successor of* is class thing. This is not problematic for books, but to categorize the Queen and former king as things is less appropriate. In Section 3.5 we further discuss the requirement for specializing data models without sacrificing interoperability, here we continue with outlining the differences between an object- and event-centric modeling approach.

3.4.2 *Challenge 2: how to choose between an event- and object-centric approach*

Two major approaches can be used to describe cultural heritage data: the event-centric and object-centric approach. The latter puts the artifact at the center of the data model. In an object-centric data model, an artifact is directly connected to the data that describes its features. An artifact has, for example, a creator, creation date, owner and location. Event-centric data models describe artifacts using related events. A definition of the class event is provided in the CIDOC-CRM documentation: *"This class comprises changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological or legal phenomena."* [15, p. 4]. A production leads to the creation of the artifact, an acquisition leads to a change of owner and a move leads to a change of location. The information that can be conveyed as well as the structure of data is impacted by a choice for one of the two approaches.

Measuring the appropriateness of a data model can be done by considering the balance between the amount of information that it is able to convey and the effort that is required to create the data structured according to the model [52]. Many cultural heritage institutions have either a collection management system or a library catalog system in place. These object-centric systems record which artifacts are part of a collection and are often the source of data published online. Attempting to convert this source data into data structured according to an event-centric data model requires much more effort than a conversion to an object-centric model. However, event-centric data could convey more detailed information about the creation, evolution and transition of artifacts over time. We regard these differences by discussing the creation of an artifact in more depth.

3.4.2.1 *Approach 2A: event-centric approach*

CIDOC-CRM is an example of a data model which uses an event-centric approach. As can be seen in Figure 10, many features of an artifact are modeled using an intermediary event. An artifact is for example related to its creator, by creating a path from artifact to actor, with a production or creation event connecting the two. For the portrait of Marten, we can now state that it was produced by a production event, in which Rembrandt was involved. Attaching attributes to this event allows us to provide more details about the creation of an artifact. We can for example state that the event took place in the year 1634 and happened in Amsterdam.

The granularity of event descriptions can be increased using composition. To illustrate, multiple events can lead to the creation of one artifact. The pistols of Napoleon exist of multiple parts, such as the barrel and the grip, which are all the result of different production

events. CIDOC-CRM caters for bundling the events leading to a creation, by decomposing an event into multiple related events using the property *consists of*. In practice, this can lead to long paths connecting an artifact to its creator: an artifact is produced by a production, which consists of a production carried out by an actor, who is identified by an appellation, which has label "Jean Le Page". Object-centric approaches are less verbose since they allow an artifact to be directly connected to a string or an agent concept with a label.

The object- and event-centric approaches can exist side by side. Most information in CIDOC-CRM is conveyed using events, but so-called shortcuts can also be used to connect instances without the use of intermediary events. The property *has current owner*, for example, connects a physical thing directly to an actor, thereby using an object-centric approach. EDM supports both the object-centric as well as the event-centric approach. The expressiveness of the event-centric constructs in EDM is however limited since it only includes the property *was present at* and class *event*. Conveying detailed event-centric information requires making these constructs more specific, as discussed in Section 3.4.1.

3.4.2.2 *Approach 2B: object-centric approach*

In EDM a stronger emphasis is given to the object-centric approach, due to the inclusion of many Dublin Core constructs. The reason for this emphasis is twofold: use of the object-centric approach is widespread and required constructs are readily available [45, p. 17]. The creator of an artifact can be indicated using the two properties *contributor* and *creator*, as shown in Figure 10. In contrast to the event-centric approach, the properties connect the artifact directly to the agent. The role of the contextual class is embedded in the properties semantics: they allow differentiating in the level of involvement of the creator. The property *creator* indicates the agent primarily responsible for the creation of the artifact, while the property *contributor* identifies someone who contributed to the artifact. Temporal information of the creation of the artifact can be conveyed using the property *created* and spatial information can be added using the property *coverage*.

Creation events bundle temporal and spatial information together with the actors involved. In contrast, the object-centric approach of Dublin Core uses three separate properties to relate an artifact to its creator, creation date and place of creation. This solution suffices if there is just one creator, but becomes problematic at the moment multiple actors with different roles are involved in the creation process. Say the barrel of the flintlock pistol was made by Jean Le Page, while Fleury Montagny engraved it. Since the date, place and creator are not connected, it is impossible to distinguish which agent was involved in what, where and when. For topics such as provenance, it is even

more important to consider multiple related events. We discuss this in more detail in the next section.

3.4.3 *Challenge 3: how to capture changes over time*

Capturing changes over time is relevant for cultural heritage data: artifacts are created, can be changed and might eventually be destroyed. There are also changes not directly affecting the artifact itself, but that for example regard ownership and location. We discuss why it is relevant to capture this temporal information using two examples: part addition and provenance. In some situations, it is useful to record if artifacts are augmented with new parts. An example of this is the changing of a frame of a painting. The portrait of Marten, for instance, has multiple fitting frames, but only one can be used at a given time. A museum needs to record which frame is in use and which other frames have been used before. This example shows that not only recording the current state of an artifact but also recording its changes is a worthwhile effort.

The provenance of an artifact is a series of events that regard the ownership of the artifact. The two pistols were for example reputedly owned by Napoleon, but after the battle of Waterloo bought by the tradesman Jean Sagermans, who gave them to his brother. Henry Sagermans, in turn, gave them to the State of the Netherlands. For many artifacts, not all provenance events are known. Tracing back owners can lead to new insights, sometimes showing that objects have been unrightfully obtained during some event, for example, the Second World War. Provenance tells something about the history of an artifact and is often highly relevant information for researchers. The ability to capture changes over time is essential to support this type of research. We discuss three approaches to record changes over time: adding textual descriptions, embedding temporal information in properties and using events.

3.4.3.1 *Approach 3A: textual descriptions of changes over time*

EDM does allow recording changes of ownership using the property *provenance*. The range of this property is a provenance statement and adding this statement as plain text adheres to the EDM guidelines [13, p. 19]. The following line is an excerpt of how the Rijksmuseum records the provenance statement of the painting *Jeremiah Lamenting the Destruction of Jerusalem* by Rembrandt: "Count Sergei Alexandrovich Stroganoff (1852-1923), St Petersburg and, after 1905, Paris; from whom, frs. 300,000, to Herman Rasch, Stockholm, 1922; from whom, fl. 150,000, to the museum, 1939". This textual description is more extensive than that of the marital portrait and the cassette of pistols since the painting was acquired during the Second World War. Although the text includes rich information, it is difficult to in-

terpret for machines and for example querying for previous owners is impossible without parsing it.

3.4.3.2 *Approach 3B: embedding temporal information in properties*

Temporal information can be embedded in the semantics of properties. The CIDOC-CRM properties *has former or current owner* and *has current owner* create a direct connection from artifact to actor. These properties embed temporal information in the properties using the words *former* and *current*. In the specification of CIDOC-CRM, it is advised to only use these properties whenever the date and place are unknown or only the current owner is known. Embedding more fine-grained temporal information into properties could be achieved by extending properties with an indicator for time, for example by creating the property *has owner in 2017*. This would, however, result in enumerating an impractical amount of properties.

As a result, object-centric approaches tend to describe one particular state of the world. In that state, the object has a specific shape, is owned by someone and is located at a place. Only considering this one state of the world refrains us from asking questions that, for example, regard changes in shape, previous owners and former locations. To illustrate, most properties of EDM originate from the object-centric Dublin Core data model. EDM has no constructs for capturing part additions or removal. As can be seen in Figure 10, there is the possibility to use the property *is part of* to record that an artifact consists out of other artifacts. This does, however, concern the current state of the artifact and not how the artifact changed over time.

3.4.3.3 *Approach 3C: recording temporal information using events*

Events can be used to record changes over time. CIDOC-CRM includes constructs that allow for fine-grained modeling of provenance and part addition and removal information. As shown in Figure 10, the model includes dedicated classes for part addition and part removal events. Adding a frame to a painting would be modeled using the following path: the frame was added by a part addition which augmented the portrait of Marten. Provenance is modeled using acquisition events. The property *changed ownership through* is used to connect an *acquisition* to the actors involved, with the properties *transferred title from* and *transferred title to*. Using this path, we can for example state that the two pistols changed ownership through an acquisition at the Oude Markt in Brussels in 1815 and thereby transferred title to Jean Sagermans.

CIDOC-CRM also includes shortcuts and extended paths for indicating the keeper of an artifact. After all, the owner and keeper of artifacts are not always the same actor. The acquisition of the marital portraits is an example: owned by the Rothschild family, both the

Louvre as well as the Rijksmuseum were interested in acquiring the works. In the end, the paintings were bought with money from the Dutch as well as the French government, who now both own half of each painting. Thus, the portrait of Marten has two owners and its keeper alternates between the Rijksmuseum and the Louvre. As may have become clear from the examples above, events allow for conveying highly detailed temporal information. However, in some cases, this extra detail can make retrieval of information more cumbersome. Accessing the object-centric *has current keeper* property would lead directly to the current keeper of an artifact. At the moment this information would be modeled using events, the keepers would have to be sorted according to date, to obtain the current keeper.

3.4.4 *Challenge 4: how to describe representations of an artifact*

Representations allow us to consider artifacts without being in the same physical space. A postcard of a statue, a poster of a painting or a recording of a concert all convey something about the represented artifact. It is important to note that by creating a representation, a new entity is created, which differs from the real-world artifact. Say a photographer takes a picture of the portrait of Marten. If we would treat the picture as the same entity as the painting, the creator of the painting would be the photographer, as well as Rembrandt. At the moment the difference between representation and artifact is not made explicit in a data model, this either leads to conflicting information or refrains us from describing the representation in more detail.

Representations can take many shapes and forms. A difference, relevant for online publication, is whether a representation is analog or digital. Analog representations are for example posters, postcards and reproductions. To illustrate, a small reproduction of The Night Watch is exhibited next to the original. The original painting by Rembrandt used to be larger, but parts of it were removed in order to fit the city hall of Amsterdam. The reproduction provides insights into how the original composition must have looked like. The Rijksmuseum rarely keeps track of analog representations, as digital representations are more important since they can accompany online descriptions of artifacts.

The range of digital representations includes images, sounds, videos and 3D models. Many types of artifacts can be represented using an image, although for artifacts such as the cassette with pistols, multiple images are required to allow inspection of all sides. Different file encodings can lead to even more representations, for example, introducing a lossless and a compressed version. For instance, the Rijksmuseum has 1083 images of the portrait of Marten alone. Many of these are close-ups of details, but also pictures taken with varying equipment, registering different light spectra, such as X-ray and

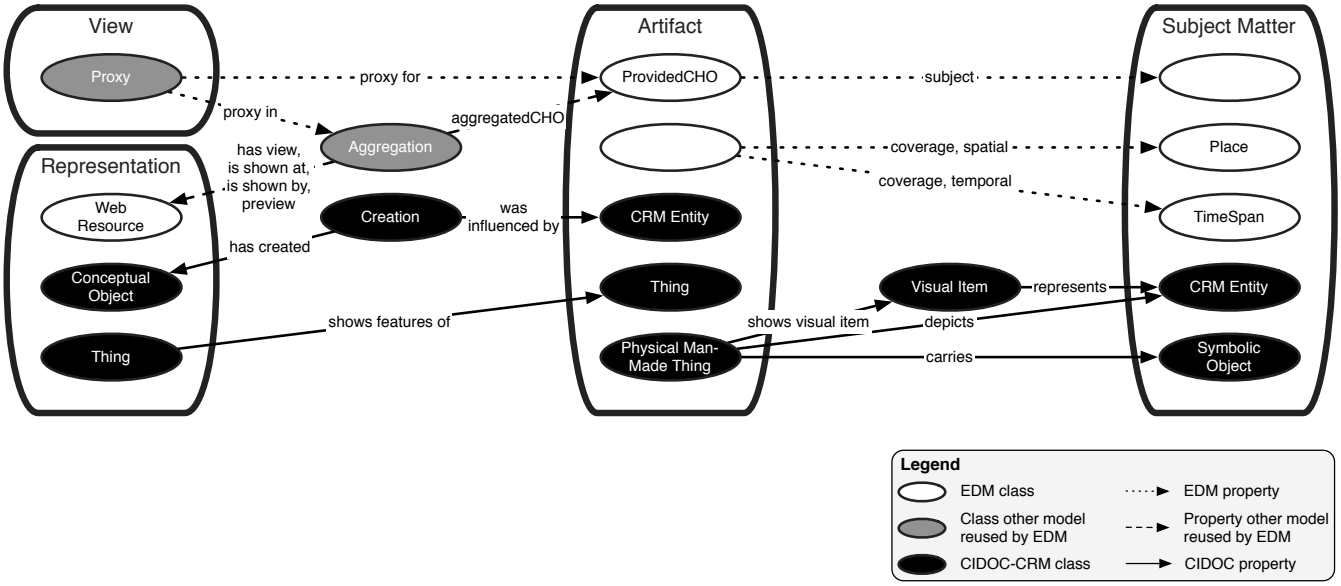

Figure 11: A representation of three key modeling aspects of publishing cultural heritage data online: subject matter, view and some representation of the artifact.

infrared. This multitude of representations makes separate descriptions of representations all the more important.

3.4.4.1 *Approach 4A: aggregations that connect artifact and representation*

EDM uses aggregations to connect data about artifacts to digital representations. An aggregation can only be connected to one artifact, but different properties can be used to connect it to multiple digital representations. The most generic property for doing so is *has view*, which does not have a range restriction beyond that the resource should be available on the web. Although this is not formally reflected in the range specification, three more specific properties limit the range of the web resource that functions as view. The range of *is shown by* is limited to digital representations in the best available quality. The property *is shown at* connects the aggregation to a website of the institution at which the artifact is shown. For the portrait of Marten this would be `https://www.rijksmuseum.nl/en/collection/SK-A-5033`. The range of *preview* is set to thumbnails that represent the artifact. Figure 11 provides an overview of these properties.

3.4.4.2 *Approach 4B: similarity of artifact and representation*

CIDOC-CRM does not have properties dedicated to connecting artifacts to representations, yet more generic properties can be used to achieve the same effect. As can be seen in Figure 11, the first approach uses a creation event, relating the representation to the artifact using

the property *was influenced by*, indicating the resemblance. A more direct connection is created with the use of the property *shows features of*, that indicates that the artifact is similar to the representation. The domain of this property should be the derivative, in this case, the digital representation. The property can be refined by adding the type of similarity. Where EDM properties can exclusively be used to refer to online representations, the properties of CIDOC-CRM can also refer to analog representations.

3.4.5 *Challenge 5: how to model multiple metadata sources with alternative views*

Metadata is a point of view: it is created by someone who describes an artifact to the best of his or her knowledge, given a certain context. The context can be different, influenced by for example the institution or the intended use of data. This makes the metadata volatile, while the physical artifact that is described is not. These differences are not problematic in an environment that considers just one context, say a collection management system keeping track of the collection objects of one museum. But at the moment data is published online, data created in different contexts starts to coexist.

This can lead to situations where data from different sources describes the same physical artifact, with potentially conflicting information. For example, the title used by the Rijksmuseum for the wedding portrait by Rembrandt is "Portret van Marten Soolmans", while the Louvre uses "Portrait de Maerten Soolmans". Besides the difference in language, the latter title is the result of the entanglement of names of Oopjen Coppits first and second husband. She married Maarten Soolmans, but after his death, she remarried Maerten Daey. For users it is essential to be aware of the context in which data is created, thereby allowing an informed decision of which information to use. It is, therefore, important that data models support capturing multiple sources describing the same artifact with possibly conflicting information.

3.4.5.1 *Approach 5A: proxies*

EDM caters for having multiple views on the same physical artifact using the class *proxy*. As shown in Figure 11, an instance of proxy is connected to two entities: it is a proxy for an aggregation and a proxy for an artifact. The aggregation bundles web resources provided by one of the institutions together. In case of the portrait, the aggregation could bundle the digital images provided by the Rijksmuseum. The artifact that is connected to the proxy is an instance of type *provided cultural heritage object*, represented by an identifier. At the moment proxies are used, data describing the artifact is not connected to the instance of *provided cultural heritage object*. Instead, the data is

connected to the proxy instance. The proxy representing the data provided by the Rijksmuseum is for example connected to "Portret van Marten Soolmans", while the proxy of the Louvre is connected to "Portrait de Maerten Soolmans".

There are two cases in which proxy constructions are useful outside the context of Europeana. The first regards other aggregators that want to be able to convey different views on the same artifact. The second concerns cultural heritage institutions that are aware that data about an artifact is already ingested by an aggregator and who want to add an additional view. The latter case will be rare, most institutions will not check for the presence of a record. As a result, an institution will not use the proxy construction at all in the data sent to an aggregator. Thus, data ingested by an aggregator needs to be manipulated to cater for multiple views [45, p. 10]. Metadata has to be moved from instances of class *provided cultural heritage object* to entities of class *proxy*.

If two institutions are describing the same physical artifact, the object identifier that the proxy refers to is ideally the same or matched. Most current identifiers redirect to locations specific to one institution. As an example, the Rijksmuseum uses `http://hdl.handle.net/10934/RM0001.COLLECT.612987` for the portrait of Marten, which points to a page of the Rijksmuseum website. It is unlikely that the Louvre will use the same identifier. So either alignment techniques have to be in place in order for the proxy construction to be useful, or a move has to be made towards institution agnostic identifiers for artifacts. The Cultural Objects Name Authority[7] which is currently under development by the Getty research institute might provide such identifiers in the future.

3.4.6 *Challenge 6: how to contextualize artifacts based on subject matter*

Properties of artifacts can be divided into perceptual information about the artifact and information on a conceptual level [41]. To illustrate, an eagle has been engraved on one of the pistols of Napoleon. While this visual item is engraved by a person on a physical man-made thing, it represents a species of birds. The portrait by Rembrandt shows the person Marten Soolmans, wearing festive clothing, including a lace collar. The portrait is however not a person nor made of lace. Therefore a clear distinction should be made between perceptual properties of the artifact (e.g. materials, dimensions) and conceptual properties (e.g. who, what, where is depicted).

Many cultural heritage artifacts convey conceptual properties in the form of subject matter: a statue might represent a person, a painting might depict an event and a book might carry a belief. The range of topics is diverse, limited only by the imagination of the creator

7 `http://www.getty.edu/research/tools/vocabularies/cona/`

and the interpretation of the beholder of the artifact. The library field uses subject matter extensively to allow retrieval of relevant books. In contrast, it is not considered part of current museum documentation practices. One of the reasons is the lack of agreement regarding the terminology that should be used [25].

For improving the accessibility of their online collection, the Rijksmuseum recognizes the value of subject matter descriptions. The museum uses the Iconclass vocabulary[8] to annotate subject matter. This vocabulary only covers a part of the topics encountered on artifacts and use of additional or more general sources of terminology are considered. Usage of terminology that crosses over institutional boundaries can prove to be a valuable point for integration of online cultural heritage data. We continue by discussing different approaches of relating artifacts to subject matter.

3.4.6.1 *Approach 6A: unbound subject matter*

Both data models include an unbound path, that allows connecting an artifact to all other entities available, illustrating the vastness of available subject matter topics. The most generic EDM subject matter property is *subject*. The range of subject is not defined and therefore all sorts of topics can be related to the artifact. The most elaborate CIDOC-CRM subject matter path connects the artifact using the property *shows visual item* to entities of class *visual item*, which in turn is connected by the property *represents* to the root of the class hierarchy of CIDOC-CRM: *CRM entity*. We can say that one of Napoleon's pistols shows the visual item of an eagle, which represents an entity of class *biological object*. The property *depicts* is a shortcut of this path, directly connecting the artifact to an entity, omitting the visual item. By connecting to the most general term, every other class in the CIDOC-CRM hierarchy can be shown on an artifact.

3.4.6.2 *Approach 6B: subject matter limited by range*

As shown in Figure 11, EDM includes three properties that limit the range of subject matter. The property *coverage* restricts the range to temporal or spatial topics. Coverage could, for example, relate the portrait by Rembrandt to Amsterdam and the second quarter of the 17th century, while subject could also relate the painting to the person Marten Soolmans. Temporal and spatial aspects can be further specified using the properties *spatial* and *temporal*. CIDOC-CRM includes a path that connects artifacts to symbolic objects using the property *carries*. This property can be used to connect non-visual works to symbolic objects, for example stating that a book carries a text. In the next section, we distill requirements from the six challenges and approaches discussed in this section. Among these requirements is the

8 http://www.iconclass.nl/

possibility to contextualize artifacts, which will become more important with the increasing amount of cultural heritage data published online.

3.5 DISCUSSION

Based upon the modeling challenges outlined in the previous section, we formulate six requirements for cultural heritage ontologies. By considering these requirements, institutions can make a more conscious choice when selecting an ontology to publish data online. The gathered requirements, in addition to the modeling approaches, are relevant and applicable to many other domains.

1. Possibility to specialize a data model without decreasing its interoperability
2. Support for recording both attributes as well as events related to objects
3. Ability to capture changes over time
4. Ability to separate descriptions of artifacts and their representations
5. Support for capturing multiple sources describing the same artifact, with possibly conflicting views
6. Possibility to contextualize artifacts using subject matter

For domains where interoperability of data is desirable, providing methods that allow integration of data from heterogeneous sources is essential. The cultural heritage domain is diverse, with institutions ranging from archives to museums. Many data models have been created, that either have specific constructs for modeling a particular type of artifact, or have generic constructs for modeling different types of artifacts. A key insight has been that generic and specific modeling approaches can be combined, by creating ontologies which can be specialized **(requirement 1)**. At the moment certain patterns reoccur often, the ontology can be extended accordingly. To illustrate, CIDOC-CRM and EDM have been refined with collection specific constructs [28, 72, 90]. Providing ontologies with a limited set of constructs, however, does put the responsibility of formulating specialized constructs at the side of institutions, while this might not be their strong suit. A danger is that institutions either solely rely on constructs provided by the ontologies, thereby losing a lot of detail contained within source data, or create flawed new constructs, that introduce inconsistencies.

The information that can be expressed is limited by the approach taken by an ontology. In the cultural heritage domain, the object-

and event-centric approaches are common **(requirement 2)**. Data published online is often the result of a conversion from an existing data source, such as a catalog or a collection management system. These sources already take a particular stance. For instance, a collection management system primarily uses attributes to describe artifacts, thereby taking an object-centric approach. Making a transition from an object- to an event-centric approach requires effort and an institution needs to assess whether this is worth the investment. Choosing an event-centric approach provides a more natural way of conveying temporal data **(requirement 3)**. Although events add a layer of complexity, the ability to capture changes over time might be vital for some use cases.

Real-world artifacts cannot be transferred over the internet, therefore we have to rely on descriptions of artifacts. A description can take the form of data, describing properties of the artifact. Representations such as images, sounds and videos can provide additional insights into the properties of an artifact. It is, however, essential to realize that a representation is a new artifact, which does not share all properties with the artifact and hence requires a separate description **(requirement 4)**. This subtle distinction becomes more apparent the moment the representation deviates more from the original. A video, like a recording of a painting from different angles, is obviously not a painting. But even "born-digital" artifacts can still have representations, such as other encodings. Both the CIDOC-CRM and EDM models include constructs for indicating how an artifact relates to a representation. These differences will become relevant in each domain that includes representations of real-world objects.

At the moment data is published online, it becomes part of an open world. This open world includes data from many sources, that possibly provide conflicting information. It is up to the user of data to decide which information to use, while understanding that data will never be one hundred percent correct and complete. Making an informed decision is enabled by the availability of the source and provenance of data. Where domain names can provide an indication of the origin of data, this context is lost when statements are made about resources outside the domain of the data publisher. For the aggregator Europeana this problem became important in an early stage since it had to manage data from many different sources, in addition to data resulting from its own enrichment strategies. EDM, therefore, provides constructs that allow for making the source of data explicit and supports having multiple descriptions of one resource **(requirement 5)**. Addressing this problem is relevant in other areas as well, for example in news stories. This will become even more important at the moment we move closer to making the internet one big data space.

Currently, the Rijksmuseum publishes Linked Data about over 350,000 objects, structured using a combination of EDM and Dublin Core. With the increase of available data, the need for contextualization rises. An important aspect of contextualization is subject matter **(requirement 6)**. Subject matter can be very diverse and often differs from the domain that cultural heritage ontologies intend to capture. Cultural heritage ontologies should, therefore, allow relating artifacts to contextual entities, even when these entities reside in different data structures.

3.6 CONCLUSION AND FUTURE WORK

The cultural heritage domain is among the first domains to embrace the Semantic Web and now features mature ontologies that can be used to structure data. In Section 3.4 we discussed modeling challenges regarding specialization, object- and event-centric approaches, temporal changes, representations, multiple views and subject matter. Considering these challenges and abstracted requirements can help cultural heritage institutions to make an informed decision about the ramifications of choosing a particular ontology. The modeling practices in addition to the gathered requirements for ontologies are relevant and applicable to many other domains.

The Rijksmuseum currently publishes information about collection objects using a combination of the EDM and Dublin Core data models. Using an event-centric model instead of the current object-centric model would overcome modeling limitations regarding changes over time and the recording of different roles of actors involved in a creation process. Although a new approach, that uses a combination of the EDM and CIDOC-CRM data models, requires a signification mapping effort, it would address all the requirements discussed in this paper. Additionally, the aim of the museum is to extend the data beyond the scope of collection management, by contextualizing objects with internal sources. These sources include bibliographic information from the art-historical library, documentation contained in the archives and research data. To convey this information adequately, ontologies from domains other than the museum sector will have to be considered.

Top-level and reference ontologies provide generic constructs, which can be refined by others. Requiring institutions to create their own specific extensions is error-prone and makes it more difficult to later align specific constructs. Creating standardized extensions for different domains and types of artifacts might help harmonize modeling efforts. This approach can already be observed within the CIDOC-CRM and EDM communities. Extensions and application profiles include models for ancient texts, fashion, archeology and scientific ob-

servations. Extending this list of topics will allow institutions to more easily publish interoperable, but detailed information.

Online publication of cultural heritage data enables the usage of data created outside the context of an institution. The increased use of controlled vocabularies in the cultural heritage sector is a first indicator of a more widespread acceptance of the usefulness of data created by others. Cross-institutional interlinking could greatly enhance the user experience, enabling a more thorough overview of the different facets of cultural heritage. At the moment more institutions are able to publish data on their own, aggregators could serve as discovery points for potential links. Increased usage of data from different sources will require data consumers to consider automated methods to validate data and assess trust in the obtained information.

The increasing amount of cultural heritage data published online will lead to new challenges. A major challenge will be ranking artifacts to adequately respond to information needs. Curators, librarians and archivist might have a natural feeling for doing this, but the required information is not always available online and with the rising number of available artifacts the need for contextualization will only grow. In the museum sector, recorded curation activities can serve as an additional source of context. However, this does not allow us to show, for example, the masterpiece of each artist in a collection. While we can provide ratings for many resources online, ranging from hotels to movies, this is rarely possible for cultural heritage artifacts. This type of subjective data is something not readily considered by cultural heritage institutions, although it might greatly improve the accessibility of information.

4

NICHESOURCING FOR CULTURAL HERITAGE

With the increase of cultural heritage data published online, the usefulness of data hinges on the quality and diversity of descriptions of collection objects. In many cases, existing descriptions are not sufficient for retrieval and research tasks, resulting in the need for more specific annotations. Eliciting such annotations is a challenge since it often requires domain-specific knowledge. Nichesourcing addresses this problem, by tapping into the expert knowledge available in niche communities. This chapter presents Accurator, a methodology for conducting nichesourcing campaigns, by addressing communities, organizing events and tailoring a web-based annotation tool to a domain of choice. We validate the methodology in three case studies, showing that it can be used to collect high-quality annotations in a variety of domains. Such annotations, in turn, can be used for search and collection integration, as described in Chapter 5 and 6.

This chapter has been submitted as "Accurator: Nichesourcing for Cultural Heritage" to the Human Computation Journal (Dijkshoorn et al. [23]) and was co-authored by Victor de Boer, Lora Aroyo and Guus Schreiber.

4.1 INTRODUCTION

Many cultural heritage collections are currently being made available online [54, 72, 90]. While such online collections can be valuable resources for the general public, scholars and professional users, their usefulness depends on correct and rich descriptions of the contained objects. Metadata describing objects is usually created by professionals working for the cultural heritage institution and typically meets the needs of other cultural heritage professionals. Many institutions lack the manpower to adapt data in order to better support different groups of users. Therefore, some institutions have turned to crowdsourcing, outsourcing tasks to a distributed and often anonymous group of people [57]. For cultural heritage organizations, crowdsourcing proved to be a low-cost solution to gather large quantities of descriptions [12, 30, 33].

While many institutions have gained significant experience with using crowdsourcing to collect large quantities of data, a remaining challenge is how to best harness the diversity in the crowd to solve difficult tasks in a sustainable fashion [55]. Describing collection objects is a knowledge-intensive task, due to the variation in types of objects, diversity in subject matter and sometimes hidden symbolic mean-

ing. Accurately annotating objects therefore often requires domain-specific knowledge. At the moment the required expertise is unavailable in an organization and when it is unfeasible to hire professionals to do the work, it is fruitful to reach out to experts within the crowd.

Nichesourcing is a type of crowdsourcing, where groups of people with domain-specific knowledge are involved in the annotation process [18, 88]. We call these groups of enthusiasts niche communities. There are numerous niche communities out there, revolving around lots of different domains. The advantages of nichesourcing are: 1) contributors are intrinsically motivated, 2) there is the potential of obtaining annotations of higher quality and 3) knowledge-intensive annotation tasks can be executed.

Where de Boer et al. [88] introduced the idea of nichesourcing and discussed small-scale case studies, a structured methodology was missing. We here present a repeatable and sustainable methodology as well as an open-source tool to support nichesourcing. The Accurator methodology and tool are developed in the context of the SEALINCMedia project[1], part of the Dutch national program COMMIT/[2]. We validate both the methodology, as well as the tool, using three extensive real-world case studies. The contribution of this chapter is fourfold:

- **Accurator nichesourcing methodology** which provides a step-by-step guide to designing and executing a nichesourcing campaign (Section 4.3)
- **Accurator annotation tool** that supports the nichesourcing process (Section 4.4)
- **Validation of nichesourcing methodology** in three case studies in different domains (Section 4.5)
- **Dataset of annotations** which includes the annotations obtained during the three campaigns (Section 4.6)

Section 4.6 includes an analysis of the annotations and an evaluation of the annotation tool. The chapter is concluded with a discussion.

4.2 RELATED WORK

Human computation is a field in which the human ability to carry out computational tasks is leveraged to solve problems that can not yet be solved by computers alone [61]. Crowdsourcing is part of the human computation field and regards tasks that are outsourced to a large group of people, often using the internet as an intermediary [24]. Crowdsourcing proved to be a good way to gather annotations at

1 http://sealincmedia.wordpress.com
2 http://commit-nl.nl

scale [2, 62]. This was recognized by the cultural heritage community and crowdsourcing has been used to annotate objects such as paintings, maps and videos [30, 32, 69]. Gathered annotations are complimentary to annotations provided by cultural heritage professionals and thereby improved the accessibility of collections [12, 33]. Crowdsourcing turned out to be a novel way of engaging the public as well [64].

Despite many successes, some crowdsourcing projects fail to live up to their expectations. Research has been conducted in classifying different types of crowdsourcing initiatives, to predict their success based on project characteristics [55]. *Methodology papers* that outline steps to successfully run a crowdsourcing campaign are however scarce. Yadav and Darlington [87] discuss guidelines to how Semantic Web technology can support the design and management of crowdsourcing projects, while Sarasua et al. [66] introduce guidelines for designing platforms hosting multiple projects. In this chapter we specify a nichesourcing methodology, contributing to the work available on crowdsourcing methodologies. More specifically, the methodology addresses crowdsourcing challenges such as solving knowledge-intensive tasks, involving experts, motivating contributors and assuring high-quality contributions.

Different approaches have been proposed to *solving knowledge-intensive, domain-specific tasks*. Ahn and Dabbish [2] introduce theme rooms, clustering tasks by domain and leaving the choice for a task to the contributor. Finding tasks can also be automated: task assignment matches characteristics of contributors with suitable tasks [14, 16]. Kulkarni et al. [51] search for experts in the crowd to improve complex, creative tasks. Combinations of improvement tasks can be optimized in crowdsourcing workflows, by considering the average ability of workers, the variance in the ability of workers and improvement difficulty [35]. Oosterman and Houben [58] invite experts from online communities to annotate objects in a specific domain. A different approach is to teach contributors how to solve knowledge-intensive tasks using a game [75]. Chamberlain [11] investigates the ability of groups on social networks to solve tasks, concluding that topic-specific groups are more active and solve more tasks. Nichesourcing builds upon these approaches by involving off- and online niche communities to solve knowledge-intensive tasks.

4.3 ACCURATOR NICHESOURCING METHODOLOGY

In this section, we describe the Accurator nichesourcing methodology. Figure 12 provides a schematic overview of the methodology, which consists of four stages: orientation, implementation, execution and evaluation. The methodology is cyclic, one iteration can build upon the results obtained during a previous iteration. The stages are

further segmented into steps. In this section, we describe for each of these steps the input, output, action and challenges. We start with a definition of nichesourcing and an introduction to the terminology used in this section.

Based upon the paper by de Boer et al. [88] we provide the following definition for nichesourcing:

> *Nichesourcing: the practice of completing knowledge-intensive tasks, by soliciting niche communities with the required domain-specific knowledge.*

Nichesourcing extends crowdsourcing in the sense that rather than executing simple micro-tasks, domain-specific, knowledge-intensive tasks can be executed by intrinsically motivated members of communities, able to provide high-quality results. Members of *niche communities* have an identity and share a domain of interest, in contrast to the crowd. These niche communities can correspond to the notion of a community of practice or interest [82]. Examples of *domains* are ornithology and fashion. We continue by listing the terminology relevant to the nichesourcing methodology:

- **requesters** initiate nichesourcing campaigns and often correspond with the cultural heritage institution that owns the collection objects.
- **collection objects** are real-world objects such as paintings or prints, of which images can be used in online applications.
- **annotations** are often short textual descriptions or concepts, that can be used to describe images of collection objects.
- **tasks** combine collection objects with the sort of annotations requested.
- **contributors** solve tasks. We refrain from using the denomination worker since there is no monetary reward given for completing tasks.
- **task difficulty** indicates how hard it is to solve a task.
- **contributor ability** is an indication of how well a contributor can solve hard tasks.

The terminology above will be used throughout the description of the different stages of the Accurator nichesourcing methodology.

4.3.1 *Orientation stage*

As shown in Figure 12, in the first stage, the goal of the campaign is determined and the objects that need to be annotated to reach this goal are identified. Based on the characteristics of these objects, the required enrichment is determined, which guides the identification of niche communities who can provide such information.

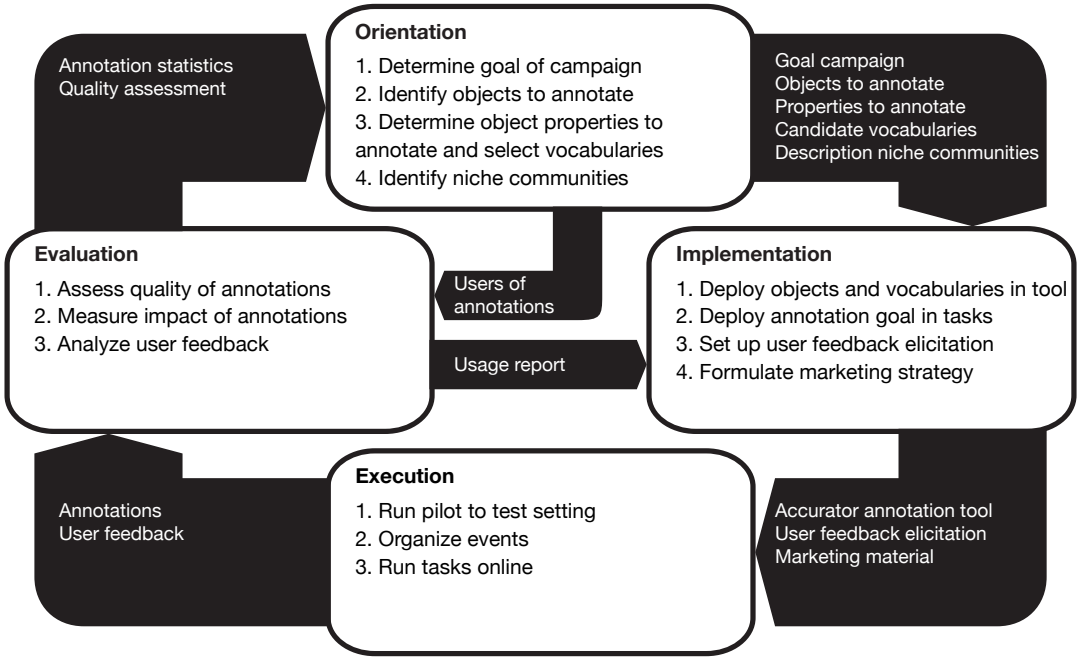

Figure 12: The four stages of the Accurator nichesourcing methodology.

1. Determine goal of campaign

input: annotation statistics *output:* goal campaign, users of annotations

action: The requester formulates a goal in this first step, stating what an institution wants to achieve with the annotations obtained during a nichesourcing campaign. The formulated goal can range from general (e.g. to improve access to a collection) to specific (e.g. to answer a digital humanities research question). A goal is based either on needs internal or external to the institution, therefore it is accompanied by an overview of the intended users of the annotations. A list of users provides clarity about who benefits from the data and gives an indication of who will take care of the collected information when the campaign is completed. At the end of the campaign, the goal is used to verify whether the gathered annotations have the desired impact. If the annotation statistics indicate the goal is reached, a new goal is formulated, otherwise, a subsequent improved campaign is used to reach the current goal.

challenges: The goal has to be formulated in such a way that contributors deem it worthy to invest time into, fitting with their domain of interest. Additionally, it helps to validate the results of a campaign if it is possible to measure whether a goal is reached. For example, if the aim is to improve access to a collection, this can be measured by standard information retrieval metrics such as precision and recall.

2. Identify objects to annotate

input: goal campaign *output:* objects to annotate

action: During this step, a subset of objects is identified, which when

correctly annotated will bring us closer to the formulated goal. This can be a) on the basis of automatic (data-driven) analysis; b) through manual selection of objects that require improvement or c) by analyzing user interactions with the collection. Most cultural heritage collections consist of objects that relate to a range of different domains. To be suitable for nichesourcing, the selected objects should share a domain, which later is matched with a community of experts.

challenges: Once a set of collection objects is either automatically or manually identified, preparation steps might be needed to ensure that basic metadata and an image are available for every object in the set. Additionally, intellectual property rights should allow images of the objects to be used online.

3. Determine object properties to annotate and select vocabularies

input: objects to annotate *output:* properties to annotate, candidate vocabularies

action: The object properties that need to be annotated to achieve the goal are identified during this step. More specifically, we distinguish properties of the objects that can be better described using numerical values, textual descriptions or concepts from structured vocabularies. One specific goal here is to identify structured vocabularies that can be used as values for the annotations. These vocabularies can be provided as input to the Accurator annotation tool, which presents concepts of the vocabulary as options to the contributors.

challenges: Cultural heritage institutions have to carefully consider which vocabularies to use for describing collection objects. The suitability of vocabularies should be assessed in terms of completeness, accuracy and original context. It can, for example, be that the vocabulary was intended to be used in a completely different context and therefore does not contain the desired concepts, or that concepts represent a worldview which is different from the institution. A lack of available labels in some language can pose a more direct problem.

4. Identify niche communities

input: quality assessment *output:* description niche communities

action: To assess the feasibility of nichesourcing, the shared domain of the set of objects should match with a niche community. These communities are identified by contemplating on which people have the expertise to annotate the objects. The characteristics of an object can for example match with professionals outside the cultural heritage sector, or with hobbyists focusing on a certain topic. A common feature of niche communities is that they can be divided into even more specialized sub-niches. It is useful to identify such sub-niches, since later in the process it helps to assign tasks to contributors most knowledgeable of a sub-niche.

challenges: The description of niche communities should include ways

of reaching out to the community, which is important for the marketing strategy in the implementation stage. Furthermore, the niche community should not only be determined on the basis of the match with the objects but more importantly, on the match with the missing information. It is not always straightforward to identify niche communities that match the selected objects and requested information. It is therefore important to allow interplay between the steps, adapting the selection or requested information to the communities available.

4.3.2 *Implementation stage*

In the second stage of the methodology shown in Figure 12, the Accurator annotation tool is deployed and tasks are designed that help reach the goal of the campaign. A marketing strategy is formulated to address the niche communities.

1. Deploy objects and vocabularies in tool

input: objects to annotate, candidate vocabularies *output:* accurator annotation tool

action: In this step, the Accurator annotation tool is deployed and relevant data is loaded. We describe the tool in more detail in Section 4.4, but in general, this requires a requester to a) set up a server environment; b) install the tool and c) adapt the tool to the domain. Once deployed, data regarding the selected objects and vocabularies is loaded. A single instance of a tool can accommodate multiple campaigns, refraining an institution from having to deploy a tool for each iteration of the methodology.

challenges: Deploying the Accurator annotation tool requires technical knowledge as well as appropriate infrastructure. Not every institution will have both readily available and therefore some might choose to outsource this step. Alternatively, an institution can choose to use existing online crowdsourcing platforms (e.g. Amazon Turk), thereby bypassing this problem. This has the downside that these platforms are not easily customized to support a particular domain.

2. Deploy annotation goal in tasks

input: goal campaign *output:* accurator annotation tool

action: During this step, the goal of the campaign is translated into smaller annotation tasks. Tasks combine objects with explicit requests for information and instructions on how this information should be provided. For a photograph, the requested information could be depicted persons, accompanied by the instruction to enter names into a text field. The identified structured vocabularies are related to requests, allowing rendering of suggestions for values to enter. Tasks are defined in the annotation tool by relating the identified objects to input fields, each accompanied by the information request and structured vocabulary.

challenges: The request for information and instructions have to be concise and unambiguous. If there is room for interpretation, this will have a negative impact on the consistency of the provided annotations. The concepts suggested can help normalize the input, but should fit the type of information requested.

3. Set up user feedback elicitation

input: - output: user feedback elicitation

action: To get insights into the behavior of users and collect feedback, user elicitation mechanisms are set up. These mechanisms can be automated and unobtrusive, such as logging interactions with the annotation tool. An institution can also choose for more direct inquiring, for example, by using questionnaires. Information gathered using these mechanisms is used to refine the orientation stage and can indicate the effectiveness of a marketing strategy. Furthermore, created user profiles can serve as input for automated quality assessment of annotations [10].

challenges: Nichesourcing relies on the intrinsic motivation of contributors. To not annoy contributors and distract them from solving tasks, the elicitation mechanisms should be as unobtrusive as possible.

4. Formulate marketing strategy

input: description niche communities *output:* marketing material and schedule

action: A marketing strategy is formulated to engage niche communities and capture the attention of contributors. This strategy includes a schedule that details when and how messages are communicated. Different outlets can be used, such as social media, newsletters and flyers. The choice of outlet depends on how the targeted niche community can best be reached. First communications are focussed on drawing attention to the campaign, by inviting people to participate in annotation events. Following an event, a message can be sent about the progress made, in addition to an invitation to keep contributing online. Subsequent communications are meant to entice people to keep participating in the campaign. At the end of the campaign, the impact of the annotations is emphasized, alongside pointing contributors towards new campaigns when available.

challenges: It can be challenging to reach the niche communities identified during the orientation stage. Sometimes organizations that already rally events around the domain of interest can serve as a point of entry. These organizations are often different from the cultural heritage institution that owns the collection. Finding a niche representative within such an organization, who is willing to collaborate, greatly eases addressing potential contributors. Another strategy is to market the nichesourcing campaign together with a broader event associated with the domain, for example, a National Week of Fashion

or an exhibition organized by the institution. This allows institutions to combine the effort needed for marketing.

4.3.3 *Execution stage*

With the tool deployed and the marketing strategy in place, the nichesourcing tasks can be executed (Figure 12). But first, tasks deployed in the annotation tool are tested during a pilot.

1. Run pilot to test the setting

input: accurator annotation tool *output:* -

action: To test the annotation tool and formulated tasks, a pilot is run with a limited number of members of the targeted niche community. During the pilot, issues are identified that should be addressed before the event. Depending on the type of issue, the subset of objects, selected vocabularies and tasks are refined.

challenges: For each issue, an assessment has to be made whether it will apply to most members of the community and therefore warrants a follow-up action.

2. Organize events

input: accurator annotation tool, user feedback elicitation, marketing material, schedule *output:* annotations, user feedback

action: Organizing an annotation event is an essential element of an Accurator nichesourcing campaign. Besides being the first source of annotations and feedback, the event is used to engage the niche community. The organization of events constitutes of three aspects: timing, location and program. With respect to timing, enough time is needed to implement the marketing strategy and advertise the event in the niche community. To make the event as attractive as possible, the event should preferably take place at a location relevant to the domain of interest. This could be at the institutions of the collection owner, or at another place relevant to the domain. The program of the event includes an introduction and demonstration of the tool. After this, contributors use the tool to annotate the collection objects. The event is concluded with a discussion, resulting in feedback which can be used during the evaluation stage. Optionally, the program can be extended with additional activities, functioning as an incentive for experts to participate.

challenges: It can be challenging to strike the right balance between time for annotating, discussion and extra activities. Enough time has to be available for annotating collection objects, in order to collect sizable amounts of annotations and to make sure that contributors have enough time to work with the tool to be able to provide feedback.

3. Run tasks online

input: accurator annotation tool, user feedback elicitation, marketing material, schedule *output:* annotations, user feedback

action: Following an annotation event, the campaign is continued online. Running the nichesourcing tasks online regards advertising the annotation tool and providing support to contributors. The interest sparked up by the event serves as initial input for advertising the tool. Updating contributors on the results of the annotation event helps to incentivize people to return at a later point in time and continuously add annotations to the collection. To sustain this attention and reach new contributors, the tool is advertised as outlined in the marketing campaign. Finding additional experts could be automated using techniques such as proposed by Kulkarni et al. [51] and Oosterman and Houben [58]. In order for contributors to not get discouraged when they run into problems, adequate support has to be available.

challenges: To sustain the interest of contributors, a cultural heritage institution will have to invest in the support and marketing of the annotation tool. When a group of contributors is actively involved in the nichesourcing campaign, the effort of marketing and providing support can be shifted towards the community [7].

4.3.4 *Evaluation stage*

At the end of the nichesourcing campaign, the impact and quality of the annotations are assessed. As shown in Figure 12, feedback gathered during the campaign is used to improve subsequent campaigns.

1. Assess quality of annotations

input: annotations *output:* quality assessment

action: The quality of annotations is assessed during this step. Quality verification procedures can be manual processes or automated processes. Both can be used within a nichesourcing campaign, although their suitability should be assessed up front. An example of a manual process is reviewing (parts of) the annotations, by contributors or professionals. An example of an automated procedure is majority voting, in which the annotation is used that most contributors added to an object. An institution decides based on the assessment, to reject or improve annotations [35]. Institutions should consider publishing the annotations along with their quality assessment since further analysis of measured disagreement can lead to new insights in crowdsourced data [44].

challenges: A relatively naive method such as majority voting might be less appropriate for nichesourcing since a small number of experts might be knowledgeable enough to provide a correct annotation. An annotation which might, in turn, contradict annotations of other contributors. Other automated approaches would, therefore, be

more suitable, for example considering trust in a contributor based on earlier annotations [10].

2. Measure impact of annotations

input: users of annotations, annotations *output:* annotation statistics

action: During this step, the verified annotations are deployed to investigate whether the goal is reached. If the goal is to improve accessibility and the user of the annotations is the institution, this, for example, entails exporting the data from the tool and incorporating the results into the collection data. At that point, a comparison of search performance of the collection with and without the annotations can provide an indication of impact [33]. If the goal cannot be reached, this evaluation serves as input for improving the next nichesourcing campaign, by for example adapting the set of objects or the properties to annotate.

challenges: Quantifying the impact of annotations can be difficult and depends on the formulated goal. Thereafter, it can be challenging to translate this evaluation towards adaptions of the next nichesourcing campaign.

3. Analyze user feedback

input: user feedback *output:* usage report

action: Feedback is gathered during events as well as online. User feedback follows from sources such as questionnaires, discussions, support requests and interaction logs. Analyzing these sources can help to improve subsequent nichesourcing campaigns. Common feedback topics regard task complexity and appropriateness of the tool. If tasks are deemed too complex, changes can be made to the selection of objects, chosen properties to annotate and the niche community which is addressed. When tasks are too easy, other crowdsourcing approaches could be considered. Feedback regarding the tool can be addressed by improving the code or choosing a different platform to deploy tasks.

challenges: Operationalizing the gathered feedback, by improving new campaigns, can be a challenge. It is, however, important to acknowledge feedback and improve the process. A contributor providing feedback took the time to work with the tool and provide feedback. If this feedback is taken seriously, a contributor might feel more inclined to contribute to a new campaign. Addressing problems with tooling requires technical skills which might not be available within an institution. The shortcomings of a tool could, therefore, be communicated to the contributors providing feedback, or programmers could be contacted to improve the tool. The Accurator annotation tool, which we discuss in the next section is open source, allowing anyone to improve the code as desired.

4.4 ACCURATOR ANNOTATION TOOL

To support the nichesourcing methodology, we present a tool called *Accurator*[3]. Accurator is a web-based annotation tool which can be instantiated for specific nichesourcing campaigns, to allow contributors to annotate images of cultural heritage objects that are automatically assigned to them. This section describes the tool and more specifically its adaptability to a specific campaign and chosen domain, as well as usability design considerations. We conclude this section by discussing how the collected data can be used by other systems and how the annotations directly impact the search functionality of the tool.

The implementation of the tool is based on Semantic Web technology [68]. The Accurator annotation tool is available as a package for the Cliopatria Semantic Web infrastructure [85]. The back-end is written in the Prolog programming language, which facilitates direct access to the data layer [83]. The front-end uses jQuery[4] and Twitter Bootstrap[5] so contributors experience an interactive and responsive tool. The source code of the package is published online, along with an in-depth guide to how new instances can be deployed[6].

4.4.1 *Adaptability to the domain*

Accurator can be customized to fit a domain, by using config files containing *domain definitions*. The definitions define links to 1) a specification of the annotation fields relevant to the domain, 2) elements of the interface tailored to the domain and 3) information that enables task assignment. Here we discuss annotation fields and interface adaption, task assignment follows in a separate subsection.

Annotation tasks are adaptable to the domain, a requester can specify field definitions for each of the annotation fields. These specifications include the field name, a short instruction and the type of field. Different types include radio buttons, check boxes and text fields. Text fields can use the auto-completion functionality, where a contributor starts to type and a drop-down menu renders alternatives related to this input, as shown in Figure 13. The contributor can either choose to annotate the object using one of these alternatives or use the entered text. The alternatives originate either from a list of values added to the field definition or from a subset of a structured vocabulary. Accurator includes Prolog predicates to identify such subsets of vocabularies, for example, based on a branch within a taxonomy.

3 http://annotate.accurator.nl/about.html
4 http://jquery.com
5 http://getbootstrap.com
6 http://github.com/rasvaan/accurator

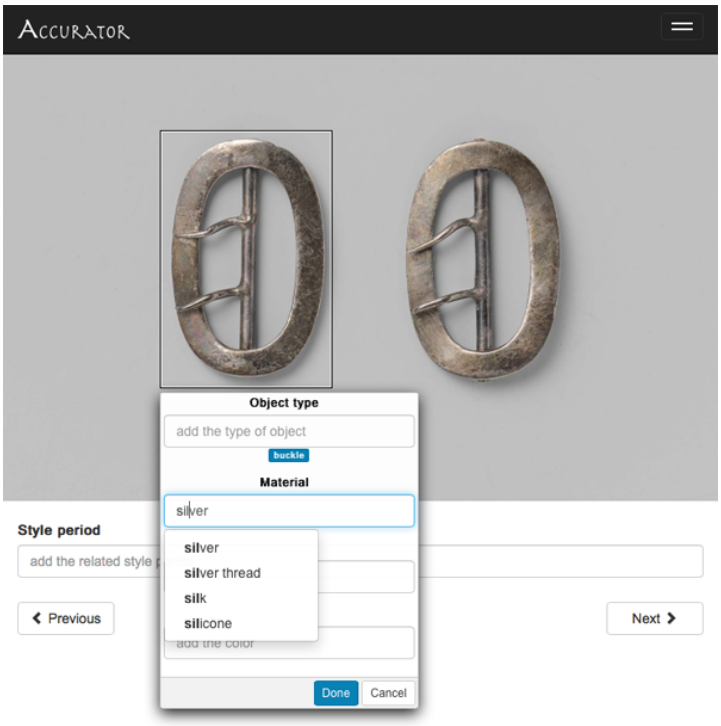

Figure 13: Annotation interface of the Accurator tool, showing the fields that can be used to annotate objects in the fashion images domain.

Annotation fields can be defined as being about the object as a whole, or be defined as being about a specific part of the object. In the first case, annotation fields are presented to a contributor alongside the image. An example of this is the style period of fashion objects. In the second case, users can draw a bounding box in the image to identify the specific part of the object that the annotation concerns, as shown in Figure 13. This allows users to annotate multiple specific elements of an object, for example, two birds of different species depicted on a print.

The default visual elements and text of the tool can be adapted as well, the default tagline used on the intro page of the tool "Help us add information to artworks" can, for example, be changed to one tailored to the fashion domain (e.g. "Help us describe fashion"). At the same time, it is possible to add images, which brand the tool with visuals related to the domain. Screenshots of tools adapted to a domain can be found in Appendix A.1, A.2 and A.3.

4.4.2 *Task assignment*

Task assignment concerns the matching of contributors with tasks. Accurator provides three modes of task assignment: *ranked*, *sub-domain based* and *recommendation*. Ranked is the default mode, which first filters out the objects already annotated by the user and sorts the re-

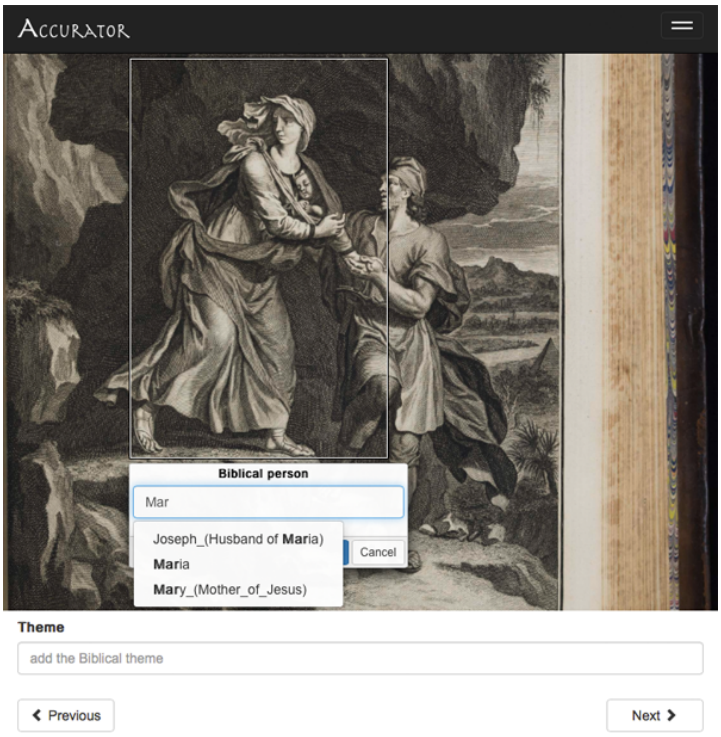

Figure 14: The annotatoin fields that can be used to annotate objects in the bible prints domain.

maining objects based on the total number of users that annotated them. A list of objects randomly picked from the least annotated objects is presented to the contributor. This is the default setting since it ensures a rapid increase in annotated objects.

In the sub-domain based mode, contributors can choose in which sub-domain they would like to annotate. To this end, a hierarchy of general and more specific domains is created, by adding references to sub-domain definitions in the configuration file of a domain definition. The fashion images domain described in Section 4.5.3 can, for example, be split into more specific domains such as costumes and jewelry, as shown in Figure 37 of Appendix A.3. The availability of sub-domains triggers a finer grained mode of task assignment, as the objects presented to contributors are filtered based on the domain they belong to. The objects from the domain chosen by the contributor are then ranked according to the ranked method described above.

The third mode of task assignment is recommendation. Recommending suitable tasks to contributors might make the annotation process more accurate and efficient. With the Accurator tool, we experimented with recommendation based on the elicitation of expertise levels of contributors. To do this, a list of expertise topics is created, the expertise levels from contributors are elicited and the obtained levels are used as input for a recommender algorithm. The list of topics is based on a structured vocabulary, referenced in the config-

uration file. In case of the birds on art domain, an example of topics can be a branch of the biological taxonomy. Contributors are asked to assess their expertise regarding each selected topic. The highest ranking topics are used as input for an explorative search algorithm, which uses the graph structure to find objects that are related to the expertise of the contributor [84]. In Section 4.6, we evaluate the three different task assignment approaches and consider the feedback of contributors.

4.4.3 *Usability*

Usability is important for crowdsourcing tools, and we argue that this is especially true for tools that are used for nichesourcing since nichesourcing relies on the intrinsic motivation of contributors. Wasting their goodwill because the tool is hard to use might make a requester miss out on valuable input. While the tool uses many Semantic Web techniques, as outlined by Sarasua et al. [66], we should not expect our contributors to be Semantic Web experts. The interface, therefore, hides technical aspects such as the persistent identifiers from contributors and uses textual labels of concepts and properties whenever available.

Part of the usability is presenting a tool in the language of the contributor. The primary language of the annotation tool is English, but many of the contributors prefer a different language. The tool supports translating textual elements of the interface, in a similar fashion as adapting texts to the domain. We translated the interface to Dutch, thereby customizing the usage for contributors from the Netherlands (Appendix A.2, Figure 35). The auto-completion alternatives are based on the labels of concepts of structured vocabularies. Oftentimes these labels are available in multiple languages. The tool is designed to render alternatives in the language of choice if available, otherwise falling back on English labels.

The Accurator annotation tool is designed to work with all regular browsers, even older versions. Therefore most contributors will be able to use the tool on their own system. The registration procedure is simple and requires minimal information to be entered by potential contributors. Questions requesting additional information about contributors used for scientific purposes are spread out over multiple blocks, each appearing after a contributor added a specified number of annotations. Additionally, system administrators are advised to use simple domain names for the online tool, so contributors can easily remember how to access the instance. We evaluated how contributors perceive the usability of the tool using a questionnaire, the results are discussed in Section 4.6.

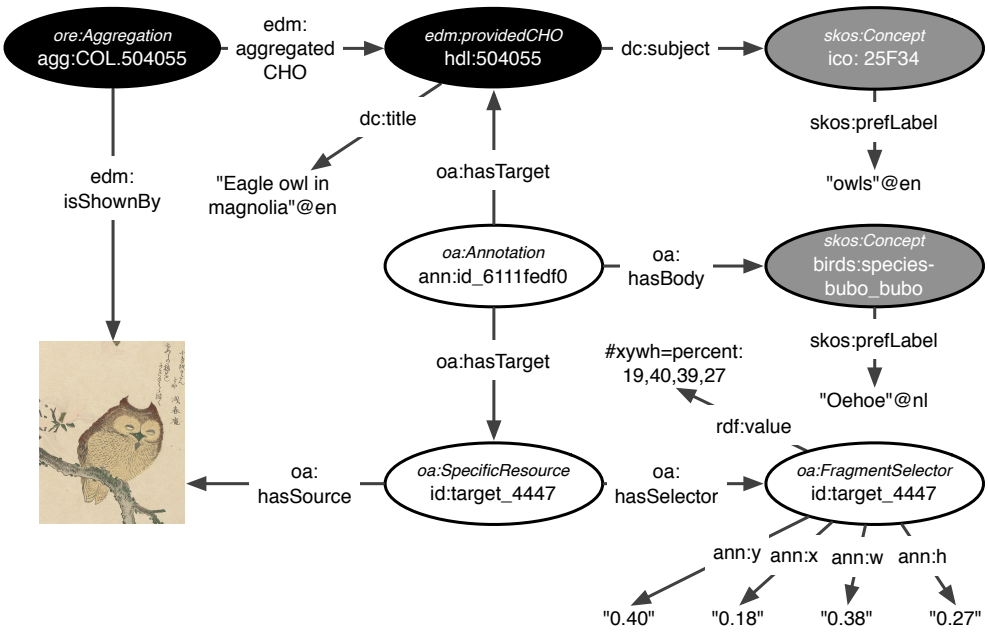

Figure 15: A graph representation of the print "Eagle owl in magnolia" annotated with the species of the depicted bird.

4.4.4 *Direct impact annotations*

Annotations added by contributors can be directly used by other systems and have a direct impact on the semantic search functionality of the Accurator tool. Annotation data is stored in the triple store of the annotation tool, which is separated from the collection management system or catalog of the institution. Using this architecture, systems that cope well with crowdsourced data can have direct access to new information, while systems relying on verified data can use exports of the information of which the quality is assessed. Storing the data using the Resource Description Framework (RDF)[7] and standardized data models improves the reusability of data. We continue by discussing the data model used within the tool, followed by a discussion of the impact of annotations on search and the ways of how the collected data can be made available.

A graph representation of information describing a print of the Rijksmuseum and an annotation acquired through Accurator is depicted in Figure 15. Constructs from the Europeana Data Model, discussed in Chapter 2, are used to model the metadata describing the object. An aggregation connects the metadata of the object with a digital representation of the object, in this case, an image. The identifier of the object is connected to metadata such as the title of the object and its subject matter. For this print, the subject matter is an Iconclass concept, representing owls.

7 http://www.w3.org/RDF/

Contributors extend the existing information by adding annotations. New annotations are modeled according to the Web Annotation Data model[8], shown as the white ovals in the figure. One annotation has as a target the object, as well as an area of the digital representation of the object. Coordinates formalizing this area correspond to the bounding box drawn by the contributor. The body of the annotation corresponds to the value selected by the contributor, in this case, a concept from the IOC bird list, with the scientific name *Bubo bubo*.

Using concepts instead of plain text to store annotations has a number of advantages. Concepts can have multiple labels, in different languages. The bird on the print of Figure 15 can, for example, be identified by its scientific name *Bubo bubo* as well as its common name in English, Eurasian eagle-owl. When a contributor enters one of these values, they refer to the same species concept. The common name in Dutch can now be used to retrieve the annotated object. This is a significant advantage over annotation using plain text since the annotations do not have to be translated every time a new language is supported. The hierarchy encountered in some vocabularies has additional benefits, for annotation and subsequent object retrieval. More general concepts can be used during annotation at the moment a contributor cannot pinpoint a specific concept. During retrieval, the tree structure can be leveraged in the other direction, if someone searches for a general concept, more specific concepts lower in the hierarchy can be included in the results as well.

Contributors can explore the collection loaded using the semantic search functionality of the Accurator tool. The search is based on a graph search algorithm, which matches keyword queries with labels in the triple store. The graph structure is used to find connected objects and clusters similar objects together [84]. Users can use this search functionality to explore the collection and find objects to annotate. Search thus functions alongside task assignment as an additional way of accessing tasks. The search algorithm is adapted to interpret added annotations as subject matter metadata, which allows users to directly inspect the result of their efforts. Observing that annotations improve the accessibility of the collection can be an added incentive to keep contributing. In Chapter 5, we analyze the impact of annotations from different vocabularies on explorative search.

The Accurator annotation tool provides multiple options to export data: annotations can be queried and exported to spreadsheets or RDF files. A public endpoint is available for queries and this can, for example, be used by systems that integrate multiple cultural heritage collections, such as the one discussed in Chapter 6. Additionally, the annotations are stored using a version management repository which can be easily published online, thereby making the results available to others.

8 http://www.w3.org/TR/annotation-model/

Table 4: Overview of the characteristics of the three case studies.

Domain	birds on art	bible prints	fashion images
Goal of campaign	improve access to collection	support comparative research	improve access to collection, investigate use of vocabularies
Objects to annotate	2,160 artworks *(Rijksmuseum)*, 406 prints *(Naturalis)*	246 bible prints *(University Library Vrije Universiteit Amsterdam)*	5,480 fashion objects *(Rijksmuseum)*
Properties to annotate & vocabularies	33,799 taxons *(IOC bird list)*, 2 genders, 3 stages of life, iconography	462 characters *(Bible ontology)*, 5,954 themes *(Iconclass)*, 34 emotions *(Emotion list)*	717 types *(Fashion thesaurus)*, 235 materials *(Fashion thesaurus)*, 117 techniques *(Fashion thesaurus)*, 20 colors *(Fashion thesaurus)*, style period
Niche community	14 bird-watchers	7 bible experts	18 fashion experts
Tool	annotate.accurator.nl	bijbel.accurator.nl	annotate.accurator.nl
Event	birdwatching event *4-10-2015 Rijksmuseum*	bible event *4-4-2016 University Library*	stitch by stitch event *23-4-2016 Rijksmuseum*
Quality assessment	comparison to gold standard	review by professional	sample review by professionals

4.5 VALIDATION OF NICHESOURCING METHODOLOGY

We validate the Accurator nichesourcing methodology using three real-world case studies in the form of nichesourcing campaigns. These show that the methodology is applicable in the highly different domains of birds on art, bible prints and fashion images. Table 4 provides a schematic overview of the cases, including links to online instances of the tool. In the following subsections, we describe each case in detail and discuss how the Accurator methodology and tool were implemented, listing the individual stages and steps of the methodology. In Section 4.6, we provide an evaluation of the quality and quantity of the resulting annotations.

Figure 16: Print by Kono Bairei, titled "Bird and red vine".

4.5.1 *Case study I: Birds on art*

The first case study regards birds depicted on objects of the Rijksmuseum Amsterdam[9]. Subject matter is diverse and sometimes outside the area of expertise of the museum's catalogers, who mostly have an art-historical background. At times this results in overly general descriptions, such as the description of the Japanese print of Figure 16: "blue-headed bird, near red vine". In 2015, the museum conducted a nichesourcing campaign, to identify birds on art. Below we show how the four stages of the Accurator methodology are applied in this case study.

ORIENTATION STAGE In collaboration with the museum, we involved experts in the process of accurately describing subject matter in order to improve access to the collection for online visitors *(determine goal)*. The first type of subject matter that the museum tried to address regarded birds. The query functionality of the museum's collection management system was used to define a set of artworks depicting birds *(identify objects)*. In this case, existing descriptions served as a sufficient basis to identify 2,160 objects. The main goal of the campaign was to accurately identify the depicted species and add this to the objects' metadata *(determine properties and select vocabularies)*. The IOC World Bird List, a comprehensive taxonomy of birds, was identified as candidate vocabulary. Other properties regarded the gender and age of the identified bird. The museum was interested whether the contributors could identify iconographic information related to the depicted birds as well.

Many bird-watchers go out into nature every week to seek birds. The museum identified them as the group of enthusiast that it was

9 http://www.rijksmuseum.nl/en

looking for *(identify niche community)*. To be able to address potential contributors within the niche community, the Naturalis Biodiversity Center[10] was contacted. This natural history museum has access to many communities, including bird-watchers. Naturalis provided an additional set of 406 prints with realistic depictions of birds, which were already annotated by the head of the vertebrate collection and could serve as gold standard for evaluation purposes.

IMPLEMENTATION STAGE The Accurator annotation tool was deployed on a server and an export of metadata of the set of objects was loaded[11], along with a conversion of the bird list[12] *(deploy tool)*. The tagline and images were changed to suit the bird domain. Screenshots of the adapted tool can be found in Appendix A.1. Short instructions for the annotation fields were written and the bird list was related to the scientific name and common name fields *(deploy tasks)*. A questionnaire inquiring about the experts' experience annotating artworks was created, to be handed out after the annotation session *(setup user feedback elicitation)*.

The campaign was marketed as "Birdwatching in the Rijksmuseum" and the event was scheduled to coincide with World Animal Day, making it easier to market *(formulate marketing strategy)*. A page was created on the museum's website[13], advertising the event and annotation tool. The biodiversity center spread the invitation to appropriate channels and the event was picked up by national broadcasters.

EXECUTION STAGE Two pilot events preceded the event, to test the stability of the system and to make employees of the biodiversity center and museum familiar with the system *(run pilot)*. The two successful pilot events resulted in small incremental updates of the system, after which the organization of the main event could start. The birdwatching event was the first event organized as part of a nichesourcing campaign and set to take place in the historical library of the Rijksmuseum *(organize events)*. To give experts an incentive to join the event, it was accompanied by various presentations related to the subject. After these talks, two and a half hours were spent annotating objects. The annotation session was closed by a curator of the museum after which people could join a bird-oriented guided tour through the museum.

Fourteen bird-watchers annotated objects during the event. Many of them brought their own books of reference (in this case bird guides) and they often formed small groups, among which tasks were discussed. For many, this was a slow paced-process, annotations were thoroughly contemplated and values for all requested data were given

10 http://www.naturalis.nl/en/

11 http://www.rijksmuseum.nl/en/api/rijksmuseum-oai-api-instructions-for-use

12 http://github.com/rasvaan/ioc

13 Website advertising the event: http://www.rijksmuseum.nl/vogelen

Figure 17: Print from Keur bible, depicting multiple biblical themes.

when possible. A flyer explained how experts could use the system at home *(run tasks online)*. Unfortunately, as part of the lessons learned, we realized it was a missed opportunity to not have advertised the online system, by sending a follow-up email to report on the results of the event and invite people to continue annotating.

EVALUATION STAGE We used the gold standard of the Naturalis-provided prints to assess annotation quality *(assess quality)*. It was not possible to feed back the results of the campaign into the collection management system of the museum, since adaptions had to be made to allow representing scientific species *(measure impact)*. Comments on the functionality of the annotation system were collected during the event and using the questionnaire. The annotations and questionnaires were analyzed *(analyze user feedback)* and in Section 4.6 we discuss the results in more detail.

4.5.2 *Case study II: Bible prints*

The second case study concerns 18th-century picture bibles. In collaboration with historians and the university library of the Vrije Universiteit Amsterdam[14], a nichesourcing campaign was conducted to enable a comparison of bibles, belonging to the Dutch Protestant heritage collection of the library. Below we describe the four stages of the nichesourcing campaign conducted in 2016.

ORIENTATION STAGE A peculiar thing about picture bibles is that a buyer could commission which prints should accompany the religious texts [70]. The prints depict bible scenes and were created by

14 http://www.ub.vu.nl

renowned artists. Figure 17 shows a digitized bible print from the collection. Analyzing which prints are included can shed light on aspects such as the popularity of artists as well as bible themes *(determine goal)*. For the historians to be able to compare bibles, the pages and the prints among them had to be annotated. The historians selected two bibles that would be interesting to compare: one printed in 1728 by de Hondt and one printed in 1729 by the brothers Keur *(identify objects)*. On request, pages of the two bibles were scanned by a company, resulting in a total of 1,003 images.

The priority of the researchers and the university library was to gather data about the subject matter of prints *(determine properties and select vocabularies)*. Two suitable structured vocabularies were identified for providing auto-completion alternatives: the bible ontology[15] is a source of biblical characters and the Iconclass vocabulary[16] includes descriptions of many biblical themes. The historians were also interested in exploring changes in emotional expressivity depicted on the prints. To allow annotation of emotions, a new vocabulary was created, based on a list of emotions of 18th-century theater texts, composed by the historians[17].

For annotating the subject matter of bible prints, an expert has to be knowledgeable about bible scripture *(identify niche community)*. The collaboration with the university library led to a fitting niche community. The library regularly organizes seminars for "friends of the university library", which often revolved around biblical topics. Since these friends of the library were willing to attend events, the library anticipated that they might also be willing to join annotation events.

IMPLEMENTATION STAGE The Accurator annotation tool was installed on a university server and customized to accommodate the bible domain. Available metadata was exported from the library catalog[18] and loaded in the annotation tool, together with the three candidate vocabularies *(deploy tool)*. Screenshots of the interface of this Accurator instance are shown in Appendix A.2. Tasks were defined by adding the fields biblical person, theme and emotion. These fields were related to parts of the candidate vocabularies and a description of the request *(deploy tasks)*. The questionnaire used for the bird domain was adapted, now inquiring about the experience of annotating biblical themes, characters and emotions *(setup user feedback elicitation)*. The library contacted the bible experts and dedicated seminars to annotation events *(formulate marketing strategy)*.

EXECUTION STAGE A pilot event was organized, during which two talks given by historians provided introductions to crowdsourc-

15 http://bibleontology.com
16 http://www.iconclass.nl
17 https://github.com/LaraHack/emotion_ontology
18 https://github.com/VUAmsterdam-UniversityLibrary/ubvu_bibles

ing and emotions in picture bibles *(run pilot)*. This did not leave enough time for an extensive annotation session, although subsequent communications with the participants led to a number of observations. Annotating bible prints required elaborate instructions about the depth and thoroughness of requested annotations. Furthermore, we observed that bible experts are not necessarily experts in recognizing depicted 18th-century emotions. Additionally, many of the digitized pages where either blank or contained only text, making it nonsensical to ask experts to annotate subject matter.

The subset selection and information to gather was adapted based on the pilot event. The task of annotating emotions was removed, to be accomplished at a later time by some other niche community. The digitized pages were classified with whether the page depicts a biblical scene. This was another annotation task but did not require expert knowledge and hence this task was accomplished using a regular crowdsourcing campaign. 246 pages depicted biblical themes and were included for the remainder of the nichesourcing campaign. In addition, a detailed step-by-step instruction manual was created to instruct people on how to use the annotation tool.

For the main annotation event, the introduction was shortened, leaving more room for annotating prints *(organize events)*. A computer room of the university library was used to host the event, with the addition of a hands-on experience with the two original historical picture bibles. Eight friends of the university library attended the annotation event and spent two and a half hours annotating bible prints. After the annotation event, participants were informed about the results of the annotation event and invited to further contribute using the online annotation tool *(run tasks online)*.

EVALUATION STAGE The annotations resulting from the events and from participants continuing at home were reviewed by library staff *(assess quality)*. Verified annotations were exported from the annotation system, published and used by the library *(measure impact)*. The library imported the annotations in its catalog, which now allows browsing based on biblical characters and themes[19]. Input for subsequent events was obtained during the event and from the questionnaires *(analyze user feedback)*.

4.5.3 *Case study III: Fashion images*

The third case study regards fashion images. In spring 2016, the Rijksmuseum organized an exhibition called Catwalk, during which fashion objects such as dresses and costumes were displayed[20]. The mu-

19 This link, for example, lists all prints with a depiction of Moses: `http://imagebase.ubvu.vu.nl/cdm/search/collection/bis/searchterm/mozes/`

20 `http://www.rijksmuseum.nl/en/catwalk`

Figure 18: Dress with train, anonymous.

seum chose this well-advertised exhibition as the context for a nichesourcing campaign.

ORIENTATION STAGE The goal of the campaign was to better describe fashion objects, thereby improving online access *(determine goal)*. Besides the museum, another user took interest in the annotations. A second goal was to support a researcher who develops a fashion thesaurus and wanted to investigate which terms contributors use to describe fashion objects. The types of objects in the fashion domain are more diverse than the prints and paintings from the previous two case studies. The museum owns a wide range of historical fashion objects, ranging from the dress depicted in Figure 18, to jewelry and prints from fashion magazines. Since the domain included such diverse types of objects, multiple subsets were identified as relevant to the fashion domain, amounting to a total of 5,480 objects *(identify objects)*.

While the objects are diverse, an information specialist of the museum determined that the information that can be gathered about the objects can be categorized under general topics *(determine properties and select vocabularies)*. These topics included technique, material, style period and color. A survey of structured vocabularies resulted in multiple possible candidates per topic, including the Art and Architecture Thesaurus (AAT) of the Getty[21]. For the campaign, it was however decided to use the fashion thesaurus created by Europeana[22], which is based on the AAT, but focusses more specifically on the fashion domain.

21 http://www.getty.edu/research/tools/vocabularies/aat/

22 http://skos.europeana.eu/api/collections/europeana:fashion.html

The diversity of the fashion domain makes it harder to pinpoint one niche community that is knowledgeable about all facets of the domain *(identify niche community)*. The community of fashion enthusiast (fashionistas) is interested in fashion in a broader sense, but they might not know much about historical objects. There are many experts working with fashion on a professional level, but describing a shoe is something completely different from describing a lace detail. Therefore, to cover as much of the diverse fashion domain as possible, the museum had to turn to a more heterogeneous group of people than the previous two case studies.

IMPLEMENTATION STAGE The collection data and structured vocabularies were loaded in the Accurator tool the Rijksmuseum already used for the birdwatching event *(deploy tool)*. For the fashion domain, six sub-domains were added: jewelry, accessories, fashion prints, paintings, costumes and lace. Contributors could choose one of these subdomains or the general fashion domain (which includes all objects of the sub-domains), to start adding annotations to. These and other interface features are shown in Appendix A.3. Similar tasks were deployed for each of these sub-domains, relating parts of the Europeana fashion thesaurus to requests for information regarding technique, material, style period and color *(deploy tasks)*.

A questionnaire focussed on the fashion domain was created to elicit feedback *(setup user feedback elicitation)*. The event was marketed using the name "Stitch by Stitch"[23] and the organization ModeMuze[24] was willing to help address niche communities *(formulate marketing strategy)*. ModeMuze is a Dutch aggregator of digitized fashion collections. This community was addressed and invited to participate in the event. Additionally, the Catwalk exhibition concluded with a conference for fashion professionals from the cultural heritage sector. An invitation was sent to these professionals as well. The event was organized following the conference, allowing professionals that attended the conference to join the annotation event.

EXECUTION STAGE Two fashion professionals participated in a small pilot, which did not bring major problems to light *(run pilot)*. The main annotation event took place in the library of the museum *(organize events)*. Since many of the contributors attended a conference in the days preceding the event, introductory talks were limited to an introduction of the annotation tool, leaving plenty of time to annotate objects. The broad invitation to different niche communities led to a diverse group of 18 contributors, including tailors, fashion curators and fashionistas. All of these contributors were asked to join

23 Website advertising the event: `http://www.rijksmuseum.nl/stitch-by-stitch`
24 `http://www.modemuze.nl`

Table 5: Results of the three case studies.

Domain	birds on art	bible prints	fashion images
Number of annotations event	835	244	1,357
Number of annotations online	307	2,138	48
Total number of annotations	1,142	2,382	1,405
Quality assessment	comparison to gold standard	review by professional	sample review by professionals
Percentage considered correct	83%	96%	84%

in a discussion at the end of the event, discussing the campaign. The annotation tool stayed online following the event *(run tasks online)*.

EVALUATION STAGE To assess the annotation quality, three fashion professionals evaluated a sample of the collected annotations *(assess quality)*. The free text annotations were compared with the structured vocabulary, seeing whether strong differences occurred, serving as input for the researcher interested in developing a fashion thesaurus for the cultural heritage domain *(measure impact)*. Furthermore, from the discussion following the event and the questionnaires filled in during the event, we received rich feedback on the annotation tool and what information about fashion can be collected *(analyze user feedback)*. The results of the analysis of the questionnaires and the annotations are given in the next section.

4.6 RESULTS

In this section, we discuss the results of the three nichesourcing campaigns and provide links to the annotation datasets. The quantity and quality of annotations provided by contributors are analyzed in Section 4.6.1. Section 4.6.2 comprises the outcomes of a user evaluation of the Accurator annotation tool, which supported the campaigns.

4.6.1 *Analysis of the annotations*

For each case study, we analyze the annotations provided in terms of two dimensions: the number of provided annotations and the quality of the annotations. For the quantitative analysis, we split the numbers by annotation field. This provides an indication of whether a field was suitable for a domain. Furthermore, each of these fields is split according to the type of input provided, differentiating between text input and the input of concepts from vocabularies. This shows whether a vocabulary covered the values adequately. The last differentiating factor is the moment the annotation was entered. This is either during an event or during a subsequent possibility to add annotations online. This allows comparing the effectiveness of the campaigns of

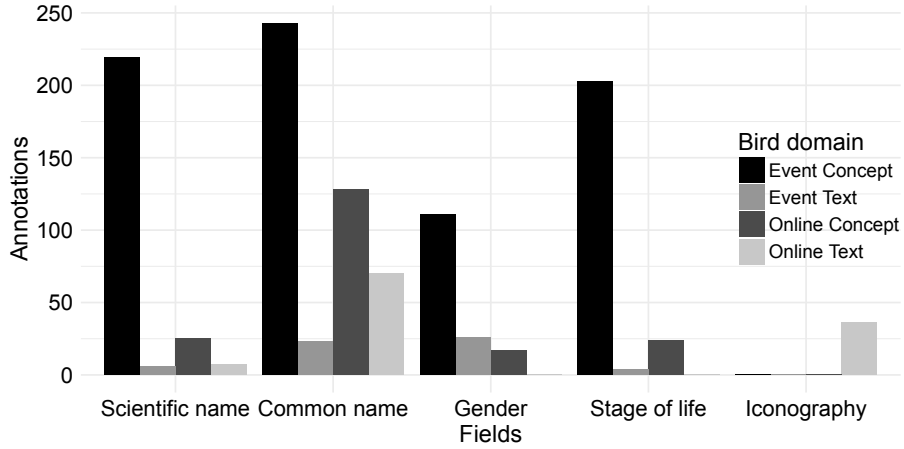

Figure 19: The number of annotations provided by contributors during the birds on art nichesourcing campaign, split by field, type of input and context of data entry.

the three case studies. Regarding the qualitative analysis, for the bird case study, a gold standard was available, allowing validation of the species annotations. The bible annotations and a subset of the fashion annotations were reviewed by professionals, providing an indication of their validity. An overview of the results is given in Table 5.

The birds on art nichesourcing campaign resulted in a total of 1,142 annotations[25], of which 835 annotations were entered during the event and 307 online. The contributors entered on average 59.6 annotations during the event. 65% of the annotations concern species and 85% of these 721 species annotations are concepts from the IOC bird list. During the annotation event, slightly more common names (266) than scientific names (225) were entered. The opposite can be observed of the annotations entered online, here there are 198 common names and 32 scientific names entered. The iconography field is rarely used: During the event, nothing was entered in this field, while online the field was used 36 times. The annotations provided during the event mainly concerned the Naturalis collection, containing prints which were not an artistic interpretation of a bird, hence the low count of iconography. A total of 231 stages of life annotations were added and 154 gender annotations. Concepts were used for the vast majority of these annotations.

The Naturalis prints allowed for evaluating the quality of the provided annotations since the depicted species were already annotated by the head of the vertebrate collection. We compare the annotations entered by contributors to this gold standard and distinguish two types of matching. The first type is a direct match of the species concept provided by the professional and the annotation of the contributor. The second type of matching concerns concepts provided by

25 Repository containing the bird annotations: `http://github.com/Rijksmuseum/accurator_annotations`

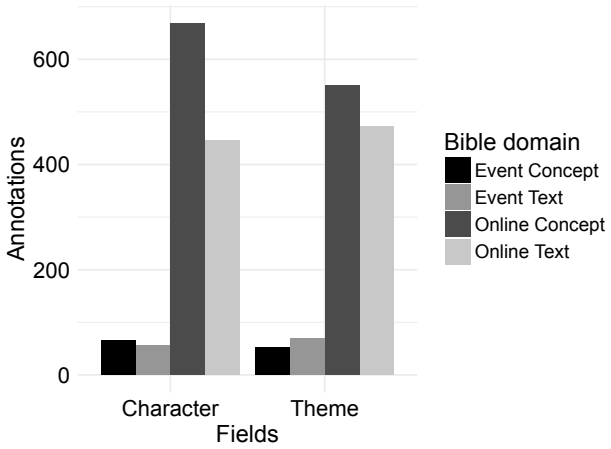

Figure 20: The number of annotations provided by contributors during the bible prints nichesourcing campaign, split by field, type of input and context of data entry.

the contributors that are one step higher up in the species taxonomy, thereby matching on a more generic level with the depicted species. Out of the 427 species annotations added to a print with gold standard, 344 of the annotations (80%) exactly match with the annotation of the professional and 11 annotations (3%) match on a more general level. This high number of correct annotations by a niche community is in-line with results observed in online groups determining species of sea slugs [11].

2,382 annotations were obtained during the bible prints campaign[26]. An overview of the obtained annotations is given in Figure 20. The event resulted in 244 annotations, a contributor added 34.9 annotations on average. In contrast to the other two domains, which have a low number of annotations added online, 90% of the bible annotations were obtained online. In total, 1,236 biblical characters were annotated, slightly more than the 1,146 themes. Vocabulary concepts were more often used than text annotations: 56% of the annotations. However, the use of concepts from structured vocabularies is lower than for example the species annotations within the bird domain. In July 2016, personnel of the university library reviewed all annotations available at that moment: 1,455 annotations in total. 96% of the annotations were accepted: 630 theme and 764 bible character annotations. These verified annotations have been added to the libraries' catalog.

A total of 1405 annotations were added during the fashion images campaign[27], as shown in Figure 21. Just 48 annotations were a result of the online campaign, 97% of all annotations were a result of the event. During this event, contributors added 75 annotations on

26 Annotation repository: `https://github.com/VUAmsterdam-UniversityLibrary/ubvu_bible_annotations`

27 Repository containing the fashion annotations: `https://github.com/Rijksmuseum/accurator_annotations`

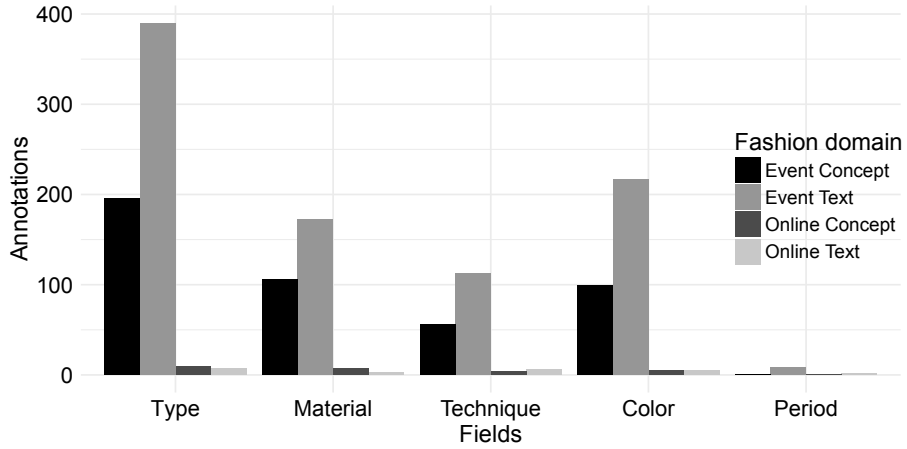

Figure 21: The number of annotations provided by contributors during the fashion images nichesourcing campaign, split by field, type of input and context of data entry.

average. 602 annotations concerned types of objects, 288 materials, 179 techniques and 326 colors. Style periods were rarely added, just 10 times. In contrast to the other two domains, the use of concepts is low: 34% of the total annotations originate out of the Europeana Fashion Thesaurus, the rest are textual annotations.

A sample of 40 annotations was evaluated by 3 fashion professionals. One professional works at the Rijksmuseum, one works for the fashion aggregator Modemuze and the last for a fashion museum in Antwerpen. Ten annotations were randomly picked from respectively the type, material, technique and color annotations. For each annotation, the professionals were asked whether it was correct, incorrect, or whether they were unable to assess it. We used majority voting to reach an assessment for 37 of the annotations, for 3 annotations the evaluations were inconclusive. Out of the sample, 89% of type annotations, 78% of the material annotation, 78% of the technique annotations and 90% of the color annotations were judged to be correct. From the in total 37 annotations upon which agreement was reached, 84% were considered correct.

4.6.2 *Evaluation Accurator annotation tool*

In order to evaluate the effectiveness of the Accurator annotation tool, at the end of each of the annotation events, questionnaires were handed out. 14 birdwatchers, 9 bible experts and 18 fashion experts filled in the questionnaire. In this section, we list the outcomes, focussing on the discussion around task assignment and the usability of the Accurator annotation tool.

During the three campaigns, different settings for task assignment were used, which are described in Section 4.4.2. The bible prints do-

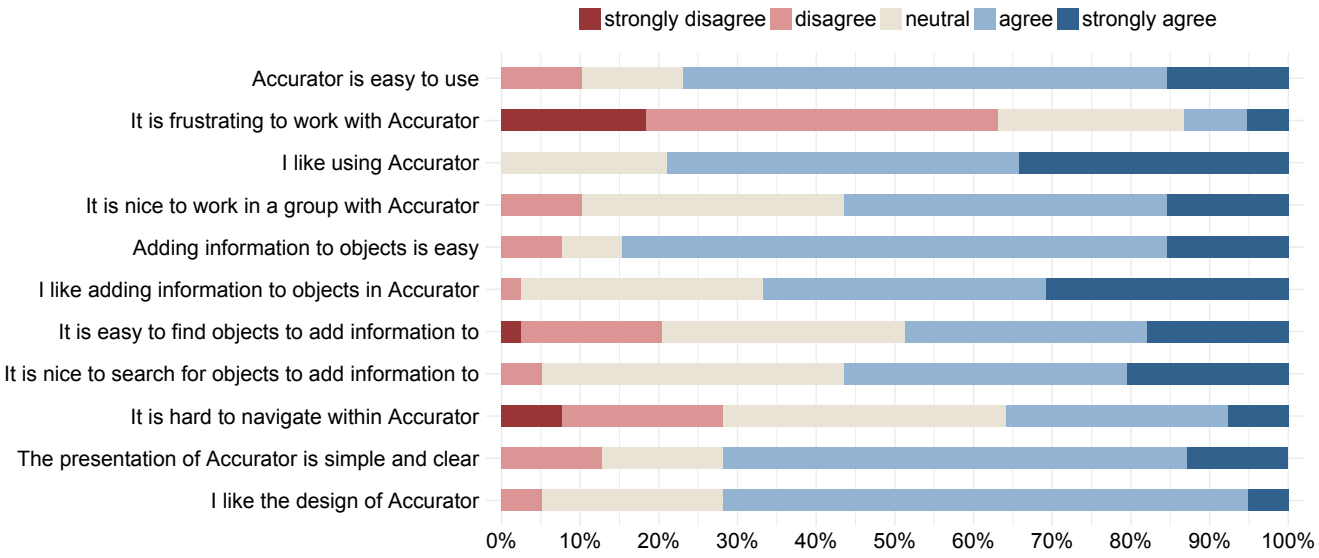

Figure 22: Overview of the answers provided that regard the usability of the Accurator annotation tool.

main used the ranked setting, the fashion images domain the sub-domain based setting and the birds on art domain used recommendation. Since the latter two are more advanced ways of assigning tasks to contributors, questions of how these settings were experienced by the contributors were included. The sub-domain based setting of the fashion images domain is deemed useful by 89% of the respondents. Contributors comment that using the sub-domains it is easier to access objects they know something about. Some would like the option to refine a sub-domain by adding filters, thereby, for example, indicating the type of accessories recommended. Additionally, dividing the domains based on style period would be appreciated. 79% of the bird watchers find recommendation based on expertise useful. They comment that it makes the process more efficient, although a different elicitation of expertise is proposed by many contributors. Expertise topics concerned different families of the biological taxonomy, while many contributors think it would be more useful to ask how much someone knows about a certain region where the birds reside.

The questionnaire included 11 statements regarding the usability of the annotation tool. Participants were asked to indicate their agreement on a five-level Likert scale, ranging from "strongly disagree" to "strongly agree". Figure 22 shows an overview of the answers of the contributors of the three case studies combined, which sums up to a total of 41 questionnaires. The evaluation of the general usability of the tool is good, 77% of the contributors agree that Accurator is easy to use. 63% disagrees with that it is frustrating to use the tool. An even higher number of contributors (79%) liked using Accurator, no one disagreed on this point. Over half of the respondents enjoyed

working with the tool in a group, although an additional one-third of the responses regards this point as neutral.

The questionnaires also included more focused questions, regarding the ability to add information to objects, searching for objects and the design of the tool. 85% of the respondents find it easy to add information to objects, two-thirds of the respondents like adding information to objects using the tool. Navigating to tasks is deemed more complicated, over half of the respondents disagreed or replied neutrally on the question whether it was easy to find an object to add information to. Over half of the contributors (56%) did enjoy searching for objects. Navigating the tool is found to be more complicated, 36% agreed that it is hard to navigate Accurator. Over two-thirds of the contributors like the design of Accurator and find the tool simple and clear.

4.7 DISCUSSION AND FUTURE WORK

Nichesourcing is a method for outsourcing tasks that require a significant level of expertise in a specific domain. The **Accurator nichesourcing methodology** presented in this chapter is geared towards executing challenging annotation tasks in the cultural heritage domain in a sustainable and repeatable fashion. The annotation events are central to the methodology and the enthusiasm with which people shared their knowledge showed the potential of this method. The **Accurator annotation tool** supports the methodology and a user evaluation indicates that the design and usability of the tool are appreciated, as well as working together with other members of the community. The **three case studies** show that the nichesourcing methodology in combination with the annotation tool can be used to collect **high-quality annotations** in a variety of domains.

While all three case studies required experts to be knowledgeable about the domain on hand, annotating fashion images proved to be the most challenging. Determining materials and techniques from single images is difficult and the formulated requests for annotations proved to be more ambiguous as well. Furthermore, the use of terms from structured vocabularies differs significantly per case study. The number of concepts used is an indicator of how suitable a vocabulary is to describe a property of a collection object. The difference in collected annotations between the event and the online tool underlines the importance of a strong marketing strategy. After the bible annotation event, multiple emails were sent inviting people to keep contributing, which clearly shows in the results.

The methodology outlined in this chapter is also **applicable outside the cultural heritage domain**. At the moment annotation tasks require expert-knowledge and niche communities with that knowledge can be identified, the nichesourcing methodology can be used.

Tasks can range from identifying species on camera trap images collected by biologists, to recognizing musical instruments in audio recording, to identifying different types of cinematography in videos. The Accurator annotation tool is just one example of the tools that can be used to collect annotations, in this case focused on images. The tool deployed in the methodology can be replaced by other tools when needed, which could, for example, be suited to annotate sound or video clips.

In **future campaigns**, we plan to optimize settings that impact the results, such as the number of selected objects, the formulation of information requests and the influence of the marketing schedule. We plan to translate the social aspect of the annotation events into the functionality of the tool and investigate whether this will retain more contributors. To accomplish goals more efficiently, we will investigate embedding nichesourcing in hybrid crowdsourcing workflows, splitting a campaign into subtasks that are solved using different methods within the human computation spectrum. This would have the benefit that for simple tasks, that can be solved by anyone in the crowd, we can resort to methods other than nichesourcing, thereby not wasting the goodwill of our expert volunteers. Other possibilities are automating parts of the campaign, such as utilizing computer vision to recognize objects on images. Finding a hybrid approach that strikes the right balance of quality and quantity of annotations will improve the usefulness of cultural heritage data published online.

5

USING LINKED DATA TO DIVERSIFY SEARCH RESULTS: A CASE STUDY IN CULTURAL HERITAGE

Large cultural heritage collections have become available online, of which the contents are unknown to the average user. Explorative search helps users to explore these collections and reach more diverse results related to their search query. In this chapter, we consider whether, and to what extent, additional semantics in the form of Linked Data can help to diversify search results. We use the Linked Data of the Rijksmuseum, extended with a number of relevant structured vocabularies. We apply an existing graph search algorithm to this data, using entries from the museum query log as the test set. Next, we analyze why some structured vocabularies have a significant effect, while others influence the results only marginally. The study shows that in this domain, search result diversity can be increased by linking collection data to structured vocabularies. This illustrates the value of enrichment strategies such as the nichesourcing methodology described in Chapter 4.

This chapter was published as "Using Linked Data to Diversify Search Results: a Case Study in Cultural Heritage" in the proceedings of the International Conference on Knowledge Engineering and Knowledge Management (Dijkshoorn et al. [19]) and was co-authored by Lora Aroyo, Guus Schreiber, Jan Wielemaker and Lizzy Jongma.

5.1 INTRODUCTION

An increasing number of large cultural heritage collections has been made accessible online[1]. Due to the sheer size of these collections, it can be challenging for users to explore them. One of the promises of Linked Data is that it can be used to improve search, by leveraging contextual information from structured vocabularies and related collections. In Chapter 4, we discussed how nichesourcing can be used to enrich metadata of objects with concepts from structured vocabularies. In this chapter, we report on an explorative search case study, in which we investigate how such enrichments influence the diversity of search results.

As data for this study, we used Linked Data of the Rijksmuseum Amsterdam (Chapter 2), enriched with a number of external vocab-

1 For example:
http://www.metmuseum.org/collections
http://www.britishmuseum.org/collection
http://www.louvre.fr/moteur-de-recherche-oeuvres

ularies that have been published as Linked Data, such as the Art & Architecture Thesaurus, WordNet and Iconclass. We employ an existing graph search algorithm to find search results [84]. This algorithm finds paths in the graph from the search query to target objects. The algorithm also clusters the results by grouping results with similar paths together. In this study, we use the number of resulting clusters and the path length as indicators of diversity. As sample queries, we collected the terms in the museum's query log for the duration of one month. We see this study as a step towards showing how Linked Data could be used to explore vast collections.

This chapter is structured as follows. In the next section, we discuss related work. Section 5.3 describes the collection data and structured vocabularies used in the study. In Section 5.4 we discuss the experimental setup, including the test set and the graph search algorithm. Results are discussed in Section 5.5. In Section 5.6 we reflect on the results and consider future directions.

5.2 RELATED WORK

A lot of work has been done on integrating cultural heritage collections and linking them to external sources. Hyvönen et al. [43] created a portal to integrated collections of Finnish museums, using Semantic Web techniques. Europeana is an initiative which supports the integration of European cultural heritage collections [47]. de Boer et al. [90] describe a methodology for publishing collections as Linked Data while preserving the rich semantics. A similar methodology is applied by Szekely et al. [72]. By integrating collections and linking them to structured vocabularies the number of available data increases, giving rise to the need for structured means to access the information [37].

Researchers at Europeana clustered artworks at different granularities, to create an overall picture and provide users with related objects [80]. The clustering approach is useful for identifying duplicate records, although at lower granularities users had difficulties explaining why artworks were clustered together. Regularities in the Linked Data cloud can also be used to cluster, with the benefit of being able to explain how objects are related. Hollink, Schreiber, and Wielinga [40] use predefined patterns to improve image search and similar paths are successfully used in a content based-recommender system [81].

There is a growing interest in the diversification of search results in the field of information retrieval. Providing more diverse results can address the ambiguity introduced by keyword queries. Agrawal et al. [1] assign topics to the user intent and documents and optimize the chance that the user is satisfied by the results. An increasing number of information retrieval systems use Linked Data to support

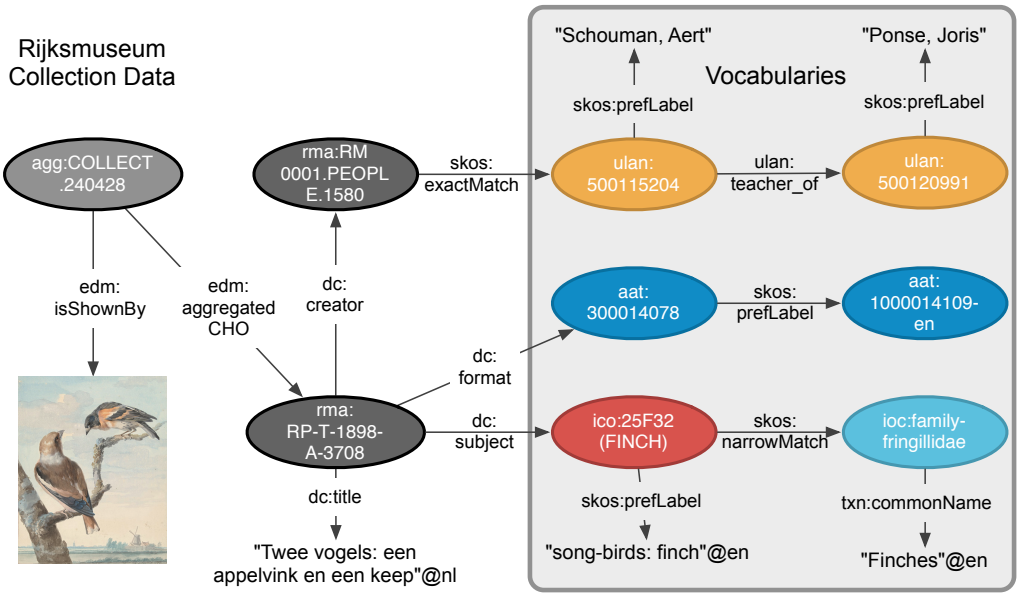

Figure 23: A graph representation of the Rijksmuseum object "Two birds", linked to four structured vocabularies.

users. Mismuseos[2] lets you search in integrated Spanish art collections and refine results using filters and facets. Constitute includes RDF representations of over 700 constitutions and lets users search and compare them[3]. Seevl uses Semantic Web techniques to provide search and discovery services over musical entities [59]. The BBC is developing a system to open up their radio archives, automatically annotating audio fragments and using crowdsourcing mechanisms to enrich the data [63].

5.3 DATA

In this section, we describe the collection data of the Rijksmuseum and the links to structured vocabularies that we used within this study. The **Rijksmuseum collection** contains over a million objects. As discussed in Chapter 2, the museum is in the process of digitizing its collection, providing access to metadata descriptions of objects using an API. The Rijksmuseum API can output data modeled according to the Europeana Data Model, a model which is more thoroughly discussed in Chapter 3. In 2014, we used this API to obtain 550,000 object descriptions.

Figure 23 shows an example of metadata of an object represented in EDM. Four pieces of metadata are shown: the title is represented as a literal, the subject is an Iconclass concept, the format points to a concept from the Art & Architect Thesaurus and the creator of the

2 http://www.mismuseos.net/
3 http://www.constituteproject.org/

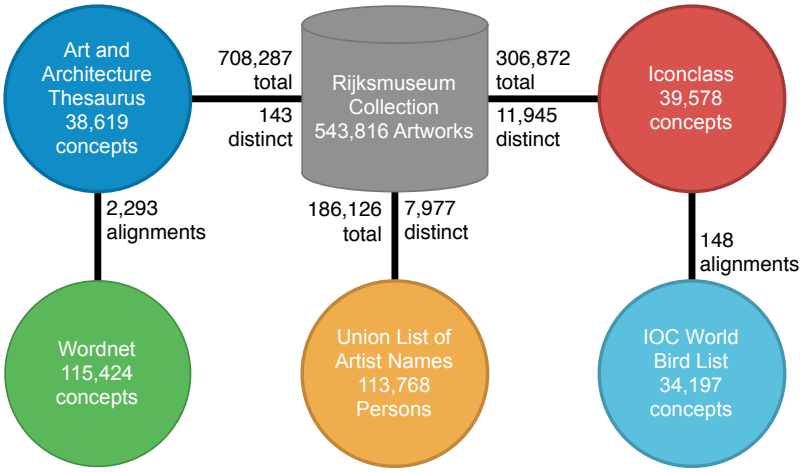

Figure 24: Overview of Rijksmuseum collection data and links to structured vocabularies.

work is represented with a resource from an in-house vocabulary of persons. The Iconclass concept is aligned with a bird concept of the IOC vocabulary. The creator resource is aligned with a corresponding resource in the Union List of Artist Names. The *teacher_of* link between the two person resources is one example of the type of extra information accessible through alignments. An overview of how the Rijksmuseum collection data is connected to structured vocabularies is given in Figure 24. Below we introduce each of these structured vocabularies more thoroughly.

The **Iconclass vocabulary**[4] is used to annotate subjects, themes and motifs in Western art. Iconclass is available as Linked Data since 2012, containing almost 40,000 concepts. Iconclass concepts are defined using a code grammar. For example, the top-level concept "Nature" has code 2; the concept "song birds" has code 25F32. The concept hierarchy is modeled using *skos:broader* and *skos:narrower* predicates. In this study, we use some 300,000 links from collection objects to Iconclass categories, created by employees of the Rijksmuseum.

The Getty research institute compiles, maintains and distributes vocabularies that focus on visual arts and architecture, in particular: 1) the **Union List of Artist Names** (ULAN), 2) the **Art & Architecture Thesaurus** (AAT), and 3) the **Thesaurus of Geographic Names** (TGN). For this study, we decided to link the Rijksmuseum collection to AAT and ULAN. AAT has 77,470 concepts describing techniques, materials and styles which artworks can have in common. In this experiment, we use the links created by the museum to the Linked Data version of AAT[5]. ULAN includes biographic information about 113,768 artists and is, in addition, a valuable source of relations between persons, such as "collaborated with" and "teacher of". We use

4 http://www.iconclass.org/
5 http://www.getty.edu/research/tools/vocabularies/lod/

the XML version of ULAN converted to RDF and create links with the collection based on string matching.

WordNet is a source of lexical information about the English language. It provides short descriptions of words, groups words with the same meaning into synsets and defines the semantic relations between those sets. We use the WordNet 2.0 version published by W3C [6], comprising over 79,000 nouns, 13,000 verbs and 3,000 adverbs. We reuse the 2,293 alignments made between AAT and WordNet by Tordai et al. [73].

The International Ornithologists Union maintains a comprehensive list of bird names. We convert the XML version of this **IOC World Bird**[7] to RDF, adding labels from the multilingual version[8]. This results in a taxonomy of 34,197 concepts describing the orders, families, genera, species and subspecies of birds and the corresponding structure. We manually align the bird concepts of Iconclass to matching concepts in the IOC vocabulary.

5.4 METHODS

5.4.1 *Experimental setup*

Firstly, we investigate how many query terms match textual descriptions in the dataset. For this purpose, we collect query terms on the Rijksmuseum website for one month (see Section 5.4.2 below). The terms are then matched with the literal index of the triple store containing the collection data and the five structured vocabularies. As the frequency of use of the query terms might be a factor that influences the number of matches, we split the list of query terms into three sublists, containing respectively the high, medium and low-frequency query terms. The query terms are split in such a way that the three sums of the number of times that the queries in a sublist are used are equal for each split.

Secondly, we explore to what extent the external semantics improve semantic search results. To this end, we use an existing semantic search algorithm (see Section 5.4.3 below for details) to perform a search on all query terms. We do this five times, each time with a different dataset configuration:

1. only collection data
2. AAT and WordNet added
3. Iconclass and IOC added
4. ULAN added
5. all vocabularies added

6 http://www.w3.org/TR/wordnet-rdf/
7 http://www.worldbirdnames.org/ioc-lists/
8 http://github.com/rasvaan/ioc

The reason for combining AAT with WordNet and Iconclass with IOC stems from the dependencies between these vocabularies, as shown in Figure 24.

The graph search algorithm delivers the results in clusters of semantically similar results. Per obtained cluster, we analyze the path length in the graph as well as the number of clusters. This gives us per query information about the average path length, average number of clusters and average number of results. The results of this analysis are again split into three parts according to the query frequencies (high, medium, low). The code developed for these experiments, as well as the resulting data, are available online[9].

5.4.2 *Query logs*

We use the query logs of January 2014 of the Rijksmuseum. From these logs, we extract all distinct query terms used, plus their frequency. This provides us with 48,733 unique query terms. We filter out 4,074 terms because they are either object identifiers[10] or are in some other way erroneous. The resulting set of 44,659 query terms is used in the experiments. The split into frequency groups of query terms results in 2,393 terms in the high split (high frequency), 16,963 query terms in the medium split (medium frequency), and 25,303 terms in the low split (low frequency).

It should be noted that these queries are made against the collection data without the structured vocabularies. This causes a bias because the collection data contains mainly Dutch terms and therefore users who have used the search interface before are likely to refrain from using English search terms, knowing that these are of limited value.

5.4.3 *Graph search*

For the experiments, we use the graph search algorithm as described in [84]. This algorithm matches the query term with literals in the triple store, using stemming. When the match exceeds a given threshold it is added to a list. The literals in this list are used as a starting point to traverse the graph-structured data. This traversal continues registering the times a specified target class is found, all the while recording the steps it makes. The starting literal and successive properties and resources used as steps form the path in the graph which serves as the basis for clustering. For clustering, the properties in the path are abstracted to their root properties when possible. In addition, resources are abstracted to their class, unless they are a concept.

9 `http://github.com/rasvaan/cluster_search_experimental_data`

10 Some problems with the existing query interface can be circumvented by entering directly an object identifier of an object, e.g. `SK-A-4979`. For the purposes of this study, we leave out the query terms resulting from this practice.

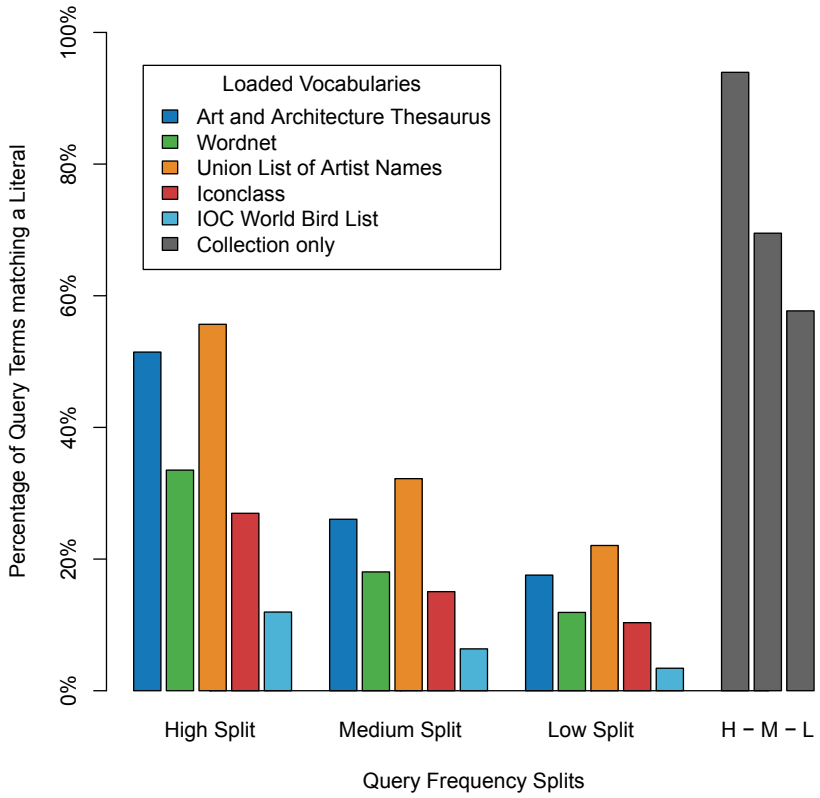

Figure 25: Bar chart showing the percentage of query terms that match with literals in the vocabularies.

This allows merging clusters based on similar semantics. The graph search algorithm is used within the Accurator annotation tool as well, as discussed in Section 4.4.4.

5.5 RESULTS

We have collected data of four types:

- The number of query terms in the test set that match text in the dataset.
- The number of search results for each of the query terms in the dataset with and without the linked vocabularies.
- The number of clusters of search results for each of the query terms in the dataset with and without the linked vocabularies.
- The distribution of path lengths of search results for each of the query terms in the dataset with and without subsets of the linked vocabularies.

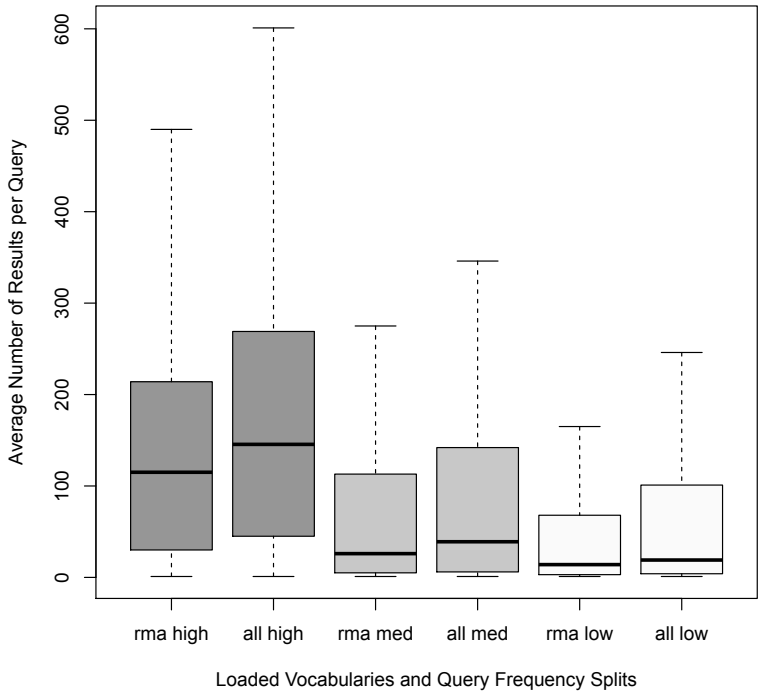

Figure 26: Overall number of search of results per query term. The boxes marked as "rma" represent the baseline (collection data only); the boxes marked with "all" represent the search results with all vocabularies loaded.

MATCHES BETWEEN QUERY TERMS AND DATASET In Figure 25 can be seen that 94% of the query terms in the high frequency split match literals in the collection data. ULAN and AAT match over 50% and continue to match many query terms in the lower splits. WordNet, Iconclass and IOC have fewer matches, with the IOC percentage on all splits below 12%. There is a decrease between the query frequency splits, were in the low split all the matches in external vocabularies are less than 23%.

To illustrate, the query term "rembrandt" matches in AAT, ULAN, and the collection data. Where in the collection and ULAN labels of "Rembrandt van Rijn" are matched, AAT matches "Rembrandt frames". The query term "watercolor" has no match in the collection data, but does match in AAT, ULAN, and WordNet. In AAT it matches materials and a technique, in ULAN descriptions of painters and in WordNet the type of paint in addition to the watercolor painting as an object. The numbers above give an indication of the potential in the data to be used for search. It depends on actual links between resources in the dataset on whether these can actually be used during search.

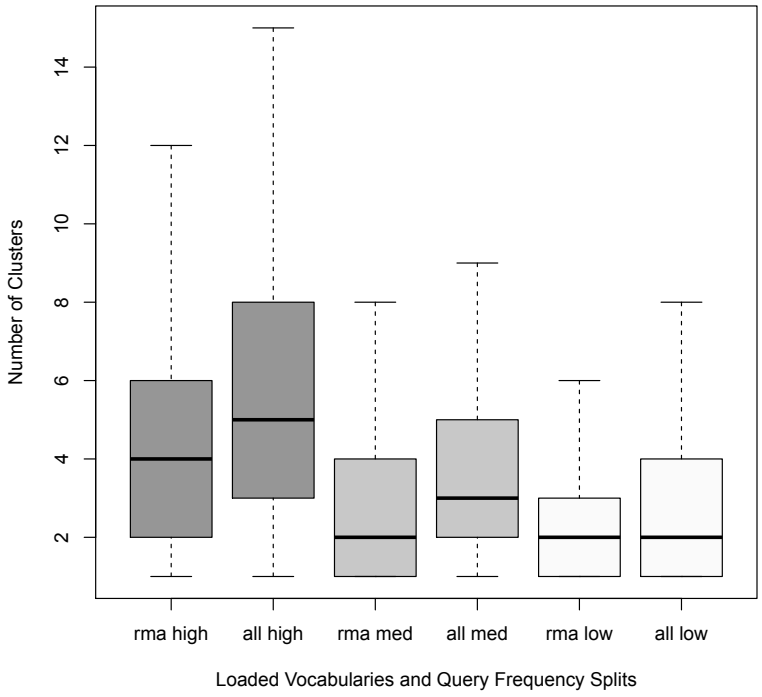

Figure 27: Box plot of the number of clusters per query. The boxes marked as "rma" represent the baseline (collection data only); the boxes marked with "all" represent the search results with all vocabularies loaded.

SEARCH RESULTS PER QUERY Figure 26 shows the overall increase of search results when the structured vocabularies are loaded. The increase is marked but moderate. The increase is highest in the third quartile of the high split; the quartile raises from 214.0 to 268.8. The mean increases from 81.5 to 104.5 search results. To give an example, when the external vocabularies are loaded, the query term "rembrandt" has 674 instead of 636 results. Instead of no results, "watercolor" increases to 9 results. It should be pointed out that the number of clusters (not shown here, see below) influences the maximum number of results, as the algorithm imposes a maximum of 100 search results per cluster.

CLUSTERS PER QUERY Figure 27 shows how the number of clusters of search results increases when the structured vocabularies are loaded. The median increases with one for the medium and high splits. There is also a marked increase in the range: some queries apparently lead to a large number of clusters. Thus, the external vocabularies not only lead to more results but also more diversified results.

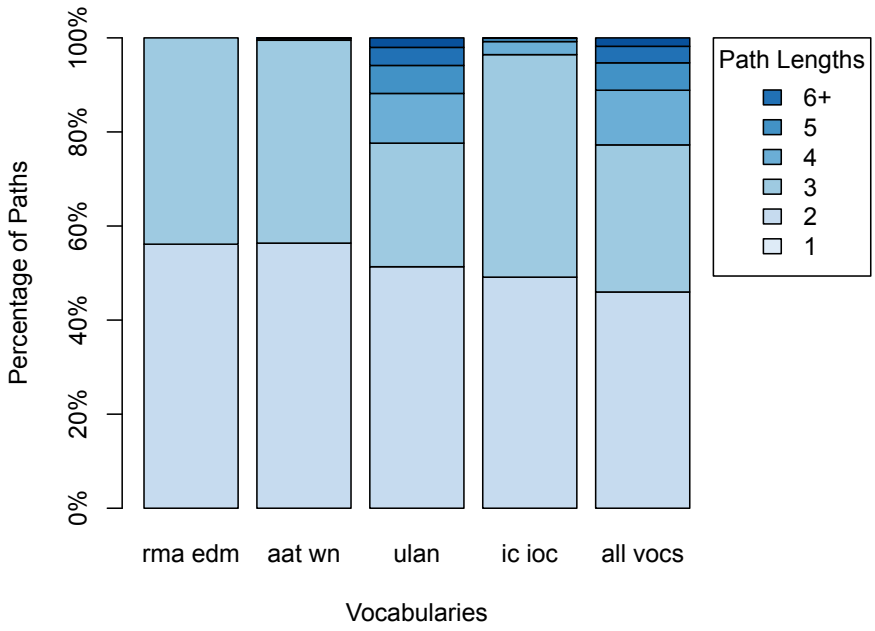

Figure 28: Path lengths of search results, shown as percentage of all search results. The first bar shows the baseline (only collection data); bars 2 to 4 depict the three (groups of) vocabularies separately; bar 5 shows the situation when all external vocabularies are loaded.

The number of clusters for the query term "rembrandt" increases from 12 to 15, adding, for example, a cluster of paintings of "Pieter Lastman" who was a teacher of Rembrandt and paintings of "Salomon Koninck" who was, according to ULAN, an ardent follower of Rembrandt. One cluster is found for "watercolor", containing watercolor paintings by "Pieter Withoos", based on the descriptive note "He specialized in watercolors of insects and flowers.". An example of a query term leading to a large number of clusters is "rubens": 8 clusters are created with the collection data loaded, 15 with the external vocabularies loaded, adding, among others, clusters about students and assistants of Peter Paul Rubens.

PATH LENGTH PER QUERY Finally, we look at the path length of search results. A longer path length suggests a diversification of results. For path length, we have looked at the contribution that the different vocabularies give to the path length. This can provide us with an indication which vocabularies are most useful.

Figure 28 shows, how the path length of the search results changes when particular vocabularies are loaded. The first bar shows the baseline, where the path length is either 1 or 2. We see that adding AAT

plus WordNet or Iconclass plus IOC has hardly any effect on the path length. The Union List of Artist Names (ULAN) has a significant effect on the path length. ULAN leads to 22% of the paths being longer than 2, up to paths of length 15.

ULAN is actually responsible for almost the complete path length diversity (see last bar). We can see an example of this phenomenon when we look again at the keyword query "rubens". The following path generated a cluster with artworks of a student of Peter Paul Rubens:

Rubens, Peter Paul → *teacher of* → *Dyck, Anthony van*
Dyck, Anthony van → *creator* → *<several artworks>*

Why does only ULAN contribute significantly to the diversity of path lengths? If we look at Figure 24, we see that ULAN has the highest number of links from collection data to distinct resources in ULAN. Also, the structure of ULAN (with many crosslinks such as "teacher" and "collaborator") makes it suitable for search diversification. In contrast, the links to AAT involve only a limited number of AAT concepts and are possibly not of much interest to users (typically things like "canvas", "oil paint" and "print").

Similar to ULAN, Iconclass has many links to collection data, in addition to a dense hierarchy of concepts. The likely reason why this does not lead to more search results is that Iconclass does not have Dutch labels. The test set of query terms came from the current search facility which works only with Dutch-language metadata. Assuming this has led to a limited usage of other languages for search, Iconclass concepts were of little use for this test set. So, this part of the results is likely to be biased by the test set.

5.6 DISCUSSION

In this chapter, we investigate whether the structure of ontologies and enrichments from structured vocabularies can improve explorative search. A more diverse palette of search results, meets the needs of users who are interested in reaching objects beyond the standard and popular ones. Moreover, diversifying search results helps institutions to promote specific, lesser-known, parts of their collection. This case study suggests that the added semantics of structured vocabularies can indeed improve explorative search.

This study does have a number of limitations. Firstly, our test set is a set of query terms that came from logs of the existing search interface of the institution involved. People, who use a search interface multiple times, are likely to limit their search to terms that work well with this interface. Therefore, the fact that the Iconclass vocabulary did not contribute a lot to search result diversity, may be a result of this bias. Secondly, there is not yet a set of standard semantic-search

algorithms. It could well be the case that other algorithms lead to different results with the same data and test set. Also, clusters and path length are indirect indicators of diversity. More studies are needed to show how valuable these indicators are and how they compare to diversity measures as introduced in [1]. Thirdly, the data set we used is limited in nature. It would be good to perform studies like these also in large, more heterogeneous datasets.

Nonetheless, this study shows the added-value of contextualization with concepts from structured vocabularies for explorative search. The results show that for this application domain we can achieve 1) an increase in the number of results, and 2) indirectly through the number of clusters and the path length, an increase in the semantic variety of search results. However, not all structured vocabularies appear to be equally useful. Based on this study we hypothesize that the usefulness of vocabularies for explorative search depends on the following two factors:

1. The number of links between distinct vocabulary resources and the metadata of target search objects
2. The richness of the internal links between vocabulary objects

It is, therefore, important that institutions consider the characteristics of vocabularies, before using them within annotation initiatives, such as the nichesourcing methodology of Chapter 4. Results clearly show that vocabularies, such as ULAN and Iconclass, which provide rich semantics for additional context (e.g. relation between people and their roles) have a significant influence on the diversity of the results. In previous studies [81] and related work [40, 80], we also show that these vocabularies are a valuable source of context and relevance for users. To conclude, additional semantics provided by structured vocabularies can help users to explore collections and reach more objects related to their search query.

6

ON THE FLY COLLECTION INTEGRATION SUPPORTED BY THE CROWD

Online cultural heritage collections often contain complementary objects, which makes integration of heterogeneous collections a worthwhile effort. In this chapter, we describe the DigiBird system that integrates four distinct nature-related cultural heritage collections. These collections either contain crowdsourced content or are enriched using crowdsourcing. Among them is a subset of the Rijksmuseum collection, enriched with annotations obtained during the birds on art nichesourcing campaign of Chapter 4. The DigiBird system illustrates the added value of these enrichments and the benefits of Linked Data for collection integration.

This chapter was published as "DigiBird: On The Fly Collection Integration Supported By The Crowd" in the proceedings of the Museums and the Web conference (Dijkshoorn et al. [20]) and was coauthored by Cristina-Iulia Bucur, Maarten Brinkerink, Sander Pieterse and Lora Aroyo.

6.1 INTRODUCTION

Collection objects from different cultural heritage institutions, when shown together, can strengthen the message conveyed individually. Traditionally this has been achieved by curating exhibitions. But, the increase of available cultural heritage data, sets a premise for collaborations that do not require physically moving the objects. Deciding on the right objects to show requires the metadata of collection objects to be adequate and interoperable. As can be read in Chapter 3, the interoperability of collection data can be improved by using ontologies.

Describing collections objects is, however, a time consuming and costly process. Crowdsourcing is evolving to be a valuable approach for cultural heritage institutions to collect metadata and engage their audience. Since the seminal Steve.Museum project [12], ample approaches have been developed in the field of crowdsourcing in the cultural heritage domain [55, 57]. The nichesourcing methodology of Chapter 4 is one of such approaches, focussed at knowledge-intensive annotation tasks. Although many crowdsourcing projects are considered successful, many initiatives still face the following four challenges:

1. Crowdsourcing initiatives are typically undertaken in isolation of other institutions and collection annotation processes

2. Crowdsourcing initiatives typically take a long time to collect the necessary annotation data

3. Crowdsourcing initiatives demand continuous promotional effort to maintain a steady user participation level

4. It is challenging for institutions to provide a structured process to incorporate the results of different crowdsourcing initiatives into their existing collection infrastructure

In this chapter, we present the results of experimenting with the DigiBird[1] system, that reinforces crowdsourcing initiatives and integrates four distinct nature-related collections by linking their crowdsourcing results. Two of these crowdsourcing platforms gather media: 1) Xeno-canto enables bird enthusiasts to collect and describe bird sounds and 2) the Dutch Species Register allows nature enthusiasts to upload images of animals encountered in the Netherlands.

The other two projects gather metadata for existing media collections: 1) Accurator, the nichesourcing methodology and annotation tool used by the Rijksmuseum Amsterdam enables birdwatchers to add bird annotations to artworks (Section 4.5.1) and 2) the Waisda? video labeling game from the Netherlands Institute of Sound and Vision allows people interested in a certain subject or domain (e.g. nature) to annotate videos related to their subject or theme of interest. The way we addressed the four crowdsourcing challenges in relation to these various collections is outlined in the remained of this chapter.

This chapter is structured as follows. In the next section, we discuss the origin of DigiBird, introducing the institutions involved. In Section 6.3 we outline how the DigiBird system addresses the four challenges listed above. The technical infrastructure of the system is described in detail in Section 6.4, focussing on data retrieval, integration and output. Section 6.5 includes examples of how the DigiBird system is used, and we conclude with a discussion of the results in Section 6.6.

6.2 ORIGIN OF DIGIBIRD

Crowdsourcing projects tend to be undertaken in isolation. This was well illustrated during the panel discussion *"Bridging the Natural Divide: Crowd-curation of Cultural Expressions Inspired by Nature"* held at MCN2014 [9]. Four crowdsourcing projects from the Netherlands were presented and each organization had developed its own system. While the initiatives all took different approaches to involve the crowd, they had one topic in common: nature. It became apparent that despite the different approaches, the collections are complementary and the profile of the user groups targeted for crowd-

1 http://www.digibird.org/

Table 6: Type of media provided by systems in addition to the crowdsourcing systems that are used to either gather collection objects or metadata describing existing objects.

	Sounds	Images	Videos
Collection		Rijksmuseum	
Crowdsourced collection	Xeno-canto	Species Register	Natuurbeelden
Crowdsourced metadata		Accurator	Waisda?

sourcing showed considerable overlap. This sparked the idea of a collaboration that extended beyond one conference panel, in which the participating institutions could explore how the different projects could strengthen one another by integrating the results. Thus DigiBird hatched.

There are two dimensions in which the crowdsourcing projects show big differences: media modality and the type of contribution by the crowd. Table 6 provides an overview of the collections and systems, mapped to these two dimensions. Appendix A.5 contains screenshots of the Species Register, Natuurbeelden, Waisda? and Xeno-canto. DigiBird includes media of three different modalities: sounds, images and videos. In the type of contributions, we distinguish how crowdsourcing is used. Some collections exist of media objects contributed by the crowd, making it a crowdsourced collection. Other systems are used to extend the metadata of existing collections. In the remainder of this section, we will discuss the individual collections and crowdsourcing methods in more depth.

The **Dutch Species Register**[2] is a thesaurus of all multicellular species observed in the Netherlands since 1758. The register is hosted by Naturalis Biodiversity Center[3] and as of January 2017, it includes 43,306 species, of which 9,644 have a corresponding image. These images are taken by amateur photographers, who upload them to the online platform. Once the images are uploaded, the depicted species and quality are validated by a group of experts coordinated by Naturalis. The register includes many images of birds, thereby making it a valuable addition to the DigiBird project. Through DigiBird, Naturalis is able to link this crowdsourced collection to both their own natural history collection and to similarly-themed cultural heritage collections from other institutions, to enrich the user experience and provide context.

Xeno-canto[4] is a foundation that aims to popularize bird sounds and recordings. An online community uploads and co-curates sounds, contributing to the goal to collect the complete sound guide of the

2 http://www.nederlandsesoorten.nl
3 http://www.naturalis.nl
4 http://www.xeno-canto.org

birds of the world. As of January 2017, a total of 9,691 species have been recorded, covering over 90 percent of all described bird species known to exist. The foundation welcomes all sorts of (re)use of their bird sound collection and they are always looking for enthusiastic people to further annotate the collection. The DigiBird system offers opportunities on both fronts.

The foundation **Natuurbeelden**[5] maintains a collection of nature videos from the Netherlands, hence the Dutch name which translates to "images from nature". The videos are shot by professional filmmakers and contributed to the collection of the foundation in a raw and uncut format. The collection of the foundation is preserved and made available by the Netherlands Institute for Sound and Vision. Sound and Vision is also involved in the development of **Waisda?**, which is an online system that allows its users to annotate audiovisual archive material in the form of a game with a purpose [33]. The goal of the game is to reach a consensus among players while they tag elements in videos. Tags are scored higher if an entry is confirmed by another player, or the tag matches a term from a controlled vocabulary. Loading the Natuurbeelden collection in Waisda? allows collecting additional metadata for the videos. The collected metadata and the DigiBird system helps Sound and Vision to explore the possibilities to take specific subject matter from their vast collection and connect them to relevant niches and other collections from other institutions.

The collection of the **Rijksmuseum**[6] includes over a million objects. As mentioned in Chapter 2, metadata and digital representations of objects are available as Linked Data. The museum realized that not all subject matter could be adequately described by its staff, since at times expert knowledge is required. Therefore, as described in Chapter 4, the **Accurator** nichesourcing methodology and annotation tool is used to gather enrichments. Accurator was developed in a collaboration of cultural heritage institutions and universities, who joined forces in the SEALINCMedia project[7], part of the COMMIT/ program[8]. Involvement of niche groups with a certain area of expertise is actively sought out by organizing nichesourcing campaigns with events tailored to specific topics. "Birds on art" is one of these topics and a birdwatching event was organized at the museum (Section 4.5.1), making the nichesourced enrichments and Rijksmuseum collection a great addition to the DigiBird system. We continue by discussing how DigiBird addresses crowdsourcing challenges.

5 http://www.natuurbeelden.nl
6 http://www.rijksmuseum.nl
7 http://sealincmedia.wordpress.com
8 http://commit-nl.nl

6.3 HOW DIGIBIRD ADDRESSES CROWDSOURCING CHALLENGES

In the introduction of this chapter, we listed four challenges for crowdsourcing initiatives. To address the first challenge, concerning the isolation of the crowdsourcing initiatives, we created the **DigiBird pipeline** that connects the above-mentioned Dutch nature crowdsourcing projects, starting with birds as a proof-of-concept. DigiBird ingests, integrates and outputs data from several systems. A central request to DigiBird is transformed into separate requests, which are delegated to the underlying systems. Obtained results from the systems are combined and returned: the data is integrated by transforming it into one representation, which can be outputted in different formats. Harmonizing data from a multitude of systems that adhere to different standards proved to be one of the most challenging tasks to address.

The second challenge, regarding the long duration of crowdsourcing data collection, was addressed by setting up centralized monitoring for the integrated systems. Since most crowdsourcing projects rely on voluntary contributions, the time it takes to collect sufficient data is unpredictable. Hence, insights into the progress of the crowdsourcing process are of great value. The DigiBird pipeline supports sending queries that retrieve aggregated statistics, such as the number of contributors and contributions. This information is shown on a **DigiBird monitoring dashboard**, tailored to each platform and updated with real-time information.

To address the third challenge, relating to the necessary continuous promotional effort of the crowdsourcing initiatives, we incorporated mechanics that trigger participants with challenging crowdsourcing tasks. We approached this by setting up crowdsourcing campaigns revolving around specific domains, inviting people to contribute by organizing events [49]. Promoting crowdsourcing initiatives is essential for keeping contributors involved. At a later stage, data collected by the DigiBird pipeline can serve as input for continuously and automatically generating crowdsourcing tasks, that incentivize contributors to keep sharing their knowledge. The **DigiBird hub** now serves as an overview of the systems, showing completed tasks.

In response to the final challenge with regard to the incorporation of crowdsourcing results into existing collections, we built the **DigiBird API** (Application Program Interface) on top of the DigiBird pipeline, which can be used by heritage institutions to embed the results of the combined crowdsourcing efforts into their online collections. The DigiBird API is already used by Naturalis Biodiversity Center to embed results on one of their sites, the Dutch Species Register. In the next section, we discuss the implementation of the DigiBird pipeline.

6.4 THE DIGIBIRD PIPELINE

It takes time to collect annotation data through crowdsourcing. Since most crowdsourcing projects in the cultural heritage domain rely on voluntary contributions, there is always a dependency on people willing to invest their time and knowledge. Within DigiBird we set three goals to reduce the time needed to obtain meaningful results from crowdsourcing efforts and make the crowdsourcing process more insightful and dynamic: harmonization of complementary collection objects, instantaneous availability of crowd contributions and the ability to monitor multiple systems in one dashboard. To achieve these goals we created a pipeline that ingests, integrates and outputs data. A detailed overview of the architecture of the DigiBird pipeline[9] is given in Figure 29. In this section we discuss the design rationale of the pipeline, emphasizing the positive and negative effects of the different ways used by institutions to make data available.

6.4.1 *Request interpretation and formulation using vocabularies*

Concepts from structured vocabularies can be of great value in determining what sort of request is made to the DigiBird pipeline and thereafter allows for correctly forwarding this request to the underlying system, through their respective APIs. This functionality is part of the request interpretation module, which disambiguates requests using the IOC World Bird List[10]. This structured vocabulary includes almost 34,000 concepts, with corresponding scientific names and labels in 23 different languages. A request can either be formulated as a common species name (e.g. "Eurasian Magpie") or a scientific name (e.g. *Pica pica*). This input is matched with a concept in the vocabulary with its corresponding identifier. Now, as the intent of the request is known, new requests have to be formulated that can be delegated to the underlying systems.

For every system, a different request has to be formulated, either in the form of a list of parameters or in the form of a query. The main source of knowledge regarding which parameters and queries can be used is the documentation of endpoints. Writing adequate documentation is a key aspect for institutions if they want developers from outside their institution to be able to work with the exposed data. In our experience, there is a lot of variation in the completeness and ease of accessing documentation, making the retrieval of data harder than it should be. Table 7 provides an overview of how four of the systems provide access to their data, whereby differences have an impact on request formulation, data retrieval and data integration.

9 http://github.com/rasvaan/digibird_api
10 http://www.worldbirdnames.org

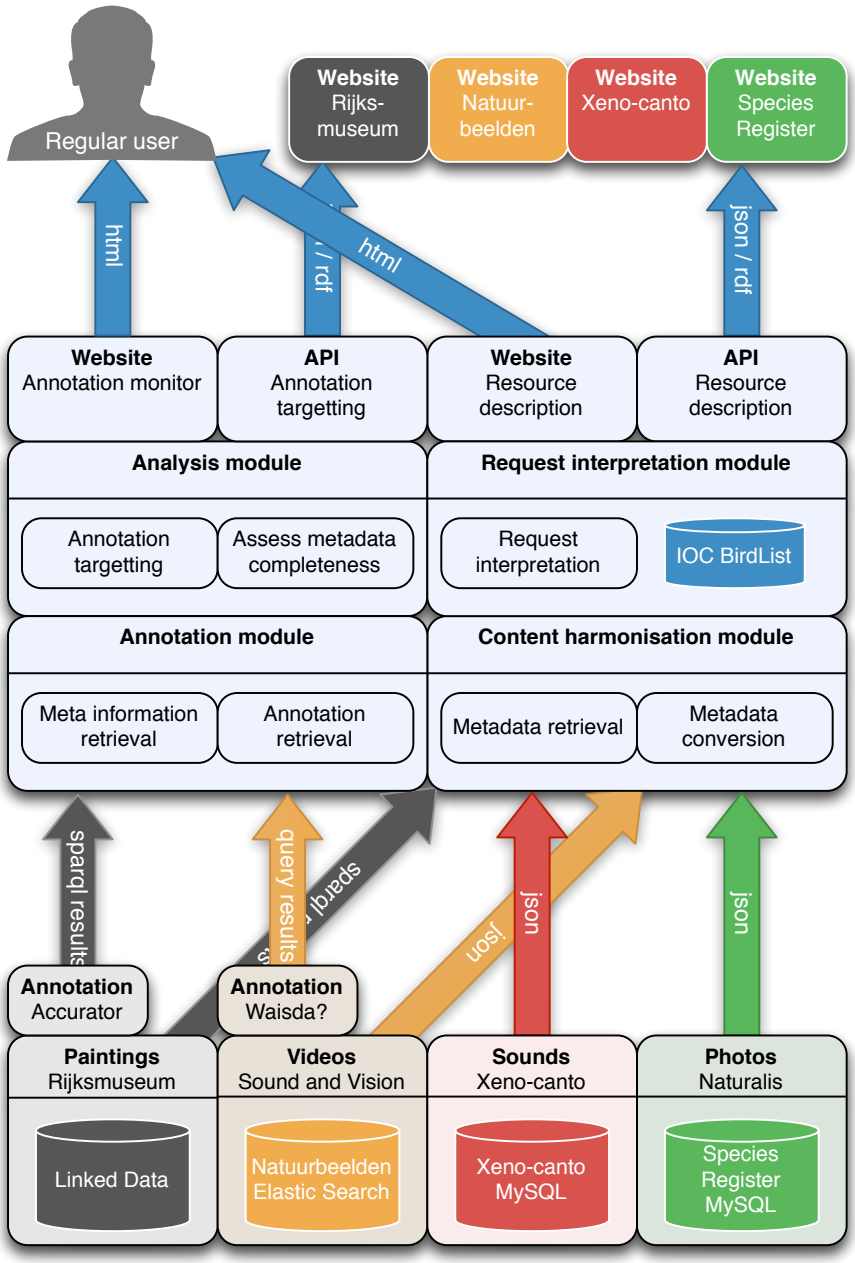

Figure 29: The architecture of the DigiBird pipeline. The pipeline ingests, integrates and outputs data obtained from collections and crowdsourcing initiatives.

Table 7: Overview of how data is obtained from four of the systems regarding the concept "Eurasian Magpie". For each institution, the supported methods of data retrieval have to be considered. The integration row refers to mapping the creator of the object to the internal data model.

Source system	Natuurbeelden	Xeno-canto	Rijksmuseum	Accurator
Type of endpoint	API	API	SPARQL	SPARQL
Retrieval method	Text search	Concept search	Text query	Concept query
Query content	Ekster	Pica pica	Ekster	ioc:Pica_pica
Response format	JSON	JSON	SPARQL	SPARQL
Metadata integration	- = -	rec = dc:creator	creator = dc:creator	creator = dc:creator

6.4.2 *Data retrieval*

An approach regularly used to obtain data for collection integration and aggregation is to download dumps of underlying databases, convert data into one format and load this in a new database [47, 53, 67]. This is a justifiable choice for stable datasets, since changes will not be missed. Another advantage of this approach is that the availability of data does not depend on other systems. However, the data used in the DigiBird project comes from dynamic systems, as a continuous stream of crowd contributions alters and extends the datasets. Since one of our goals is to make crowd contributions instantaneously available, we have to directly access underlying systems to be able to incorporate updated content immediately.

Relevant for data retrieval are the differences in types of endpoints and supported methods for matching objects to a request. A common approach is to provide an Application Program Interface (API), which specifies a set of actions that an application can undertake to interact with data. The API providing access to the metadata of Stichting Natuurbeelden does not include scientific names and is formulated in Dutch. This means that we have to rely on **text search** of descriptions in combination with the Dutch label of the concept, in this case, "Ekster" for "Eurasian Magpie".

Xeno-canto also uses an API[11] to provide access to their data, but supports parameters stating the genus and species. These two param-

11 http://www.xeno-canto.org/article/153

eters allow us to **search for scientific names**, that are present in the metadata of collection objects. This allows for far more specific retrieval of objects related to the user query, as now we can for example directly query for *Pica pica*.

The Rijksmuseum dataset and enrichments gathered with the Accurator annotation tool are accessible through a public query endpoint that supports SPARQL queries (Chapter 2 and 4). Although the query endpoints and underlying data models are similar, there is still a difference in how queries can be formulated. The Rijksmuseum objects are not linked to concepts from the IOC vocabulary. To retrieve relevant objects from the collection we use a **text query**, that selects objects that include the common name in their description. This method is prone to ambiguity issues, since some of the names of birds also refer to other types of concepts in descriptions. For example, the Dutch name for the "Great Bustard" is "grote trap", which translates to "big staircase", thus when performing a query for such a bird, one might retrieve objects related to staircases.

Using the Accurator annotation tool, the objects are linked to concepts from the IOC vocabulary, allowing a completely different **concept query**. This way, a user can search directly for objects annotated with the concepts originating from the structured vocabulary. In Accurator we can, for example, query for the concept with the URI `http://purl.org/vocab/ioc/species-pica_pica`, which refers to the "Eurasian Magpie". These two different types of queries have a big impact on performance, as querying for concepts is faster than the text queries.

6.4.3 *Data integration*

To achieve the goal of integrating complementary results, we convert the data retrieved from institutions into one internal data model. This entails dealing with the different metadata formats: not all collections use standardized data models and even the ones that do might use different standards. An alternative choice would be to directly transmit the obtained data without converting it, leaving the problem at the side of the party requesting the data. This would be inconsistent with our goal to create harmonized data and restrains us from outputting the results in different serializations. We, therefore, chose to convert the obtained data to a standardized data model. For every collection, we have to analyze which elements we are representing and how we can model these elements correctly. We outline our rationale for using elements of the Europeana Data Model[12] as our data model.

The Rijksmuseum has metadata about objects that we want to include in our results. Basic metadata includes the title and creator of

12 `http://pro.europeana.eu/edm-documentation`

a work, for example, the painting *The Contemplative Magpie*, created by Melchior d'Hondecoeter, around the year 1678. These three pieces of information can be modeled using the properties *dcterms:title*, *dcterms:creator* and *dcterms:created* from the Dublin Core Metadata Initiative[13]. This information would be sufficient if we would only be interested in the real-world object. However, we also want to communicate information about an image of this artwork.

The image has different metadata, *dcterms:creator* corresponds to the photographer who took the image and *dcterms:created* to the date the image was taken. As discussed in Section 3.4.4.1, the Europeana Data Model uses properties and classes based upon the Object Reuse and Exchange data model[14] to make a distinction between a real-world object and its digital representation. An aggregation object connects the metadata of the cultural heritage object using the property *edm:aggregatedCHO* and its digital representation using the property *edm:hasView*. The Rijksmuseum collection is the most conventional online cultural heritage collection that we use in the DigiBird system and modeling the other collections brings new modeling challenges.

Xeno-canto collects recordings of bird sounds, which leads us to ask the conceptual question: *is the primary entity the sound or is it the recording of the sound?* The sound itself is not persistent, it occurred at a certain point in time and space, which we can describe as an event. The creator of the sound is the bird and the recording is a representation of possibly multiple bird sounds. Additionally, consider as an example the analogy of a still life of a vase containing flowers. Is this a representation of a flower created by the painter, making the flower the cultural heritage object, or is the still life the cultural heritage object with the flower as subject? For the DigiBird system, we consider the recording the primary entity. This makes the recordist the creator and allows us to still capture metadata about the time and place the sound was recorded, while the bird is the subject of the recording. A particular view of the recording is the sound file hosted by Xeno-canto.

The Dutch Species Register documents all species in the Netherlands and includes images of specimens. The register is structured according to the biology taxonomy. A modeling choice would be to consider a species to be the primary entity, making images of the species views. This choice, however, leads to overgeneralization, since the images are taken at a specific point in time and space, while they show specimens of the species. If we would take the species as the primary entity, its creator and creation date would be hard to pinpoint. For the DigiBird system, we instead chose the image to be our primary entity and we can record the metadata of the image.

13 http://dublincore.org/

14 https://www.openarchives.org/ore/

6.4.4 *Data output*

The data outputted by the DigiBird pipeline can be divided into three categories: objects, annotations and aggregate information. The discussed pipeline as described up to now concerned the retrieval and integration of objects. Annotations are retrieved in a similar fashion and extend object data. Similar to the data structure outlined in Section 4.4.4, an annotation is structured according to the Web Annotation Model[15] and includes information regarding its creator and its creation date. The annotations are sorted by creation date and can be limited by providing a date range.

For the monitoring of progress of crowdsourcing systems, we are interested in a different level of the data: aggregate information regarding the number of objects, contributions and contributors. The pipeline to obtain this information is similar to obtaining objects. For the systems with a public endpoint, it is possible to write a count query obtaining the aforementioned information. For the systems offering an API, separate requests have to be available for obtaining this information. Some of the systems already supported requests for aggregate information, but for others, these had to be created.

The outputted format of the data is determined by the type of request. Three options can be used for formatting the data, requesting it as RDF, HTML or as a JSON reply. If someone browses to a DigiBird URL and provides the name of a species[16], objects related to that species are returned. If a developer provides the same species but adds in the accept header of the request that this should be in a different format, the DigiBird system supports this. It is possible to output the internal representation of data into different formats. In the next section, we discuss examples of using the data.

6.5 USING CROWD CONTRIBUTIONS AND INTEGRATED RESULTS

For cultural heritage institutions, it is often a hurdle to incorporate the results of crowdsourcing projects into the institution's existing digital infrastructure. There are two main reasons for this. Firstly, crowdsourcing projects tend to be short-lived; they run for a limited time, with limited resources, often in a research or "pilot" context. Secondly, they are often hosted on specialized platforms that are not directly connected to the core infrastructure of the institution. Examples of specialized crowdsourcing platforms born and used in the heritage sector include Many Hands [55], Zooniverse[17], Steve.Museum, but also the Accurator and Waisda? platforms used in DigiBird. After a crowdsourcing project ends, the question often arises how to integrate

15 http://www.w3.org/TR/annotation-model
16 For example: http://www.digibird.org/species?genus=pica&species=pica
17 http://www.zooniverse.org

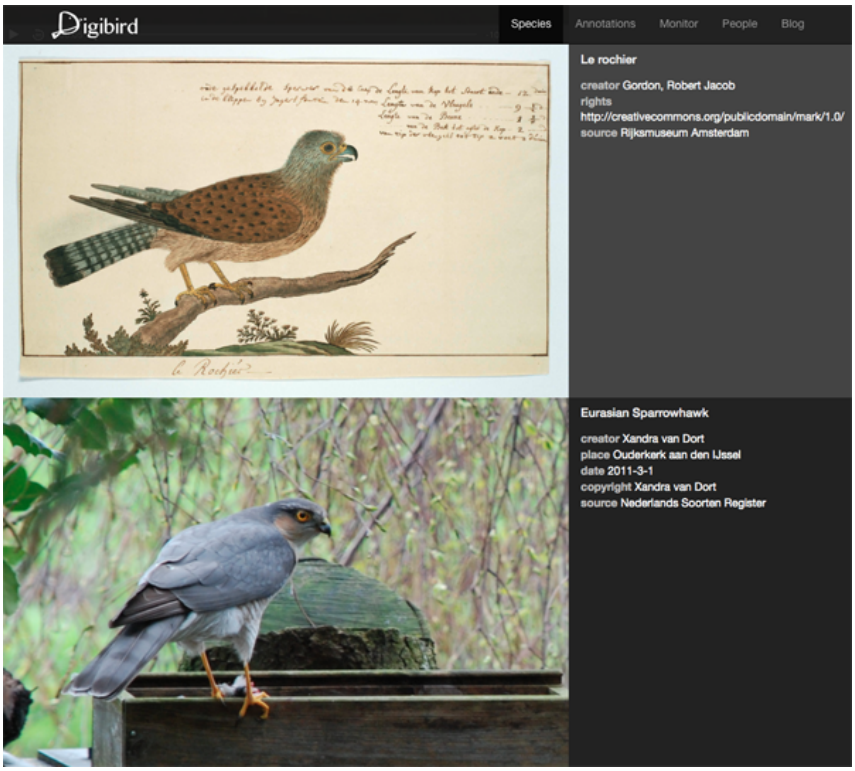

Figure 30: Screenshot of the search results page of DigiBird, showing results for the query "Eurasian Sparrowhawk" from different institutions.

or connect the results - tags, annotations, user-generated multimedia, etc. - into the collection management systems and online collection portals.

To overcome this challenge for the crowdsourcing platforms of DigiBird, we chose to connect these platforms by building an API on top of the DigiBird pipeline[18]. This API can be used by heritage institutions to embed the results of the combined crowdsourcing project results directly into their online collections. This way, institutions can connect the results of their own crowdsourcing projects to their online collections in one platform (e.g. connect Waisda? results to Natuurbeelden or Accurator results to the Rijksmuseum collection). Moreover, institutions can embed relevant results from other crowdsourcing projects in their own online collection portals as well, causing real-time cross-pollination of projects otherwise undertaken in isolation.

To showcase the functionality of the DigiBird pipeline we build a demonstrator[19] with an annotation wall and species search interface[20]. Figure 30 shows a screenshot of the search interface where a user can enter a bird species, which is auto-completed with resources

18 http://github.com/rasvaan/digibird_api
19 http://www.digibird.org
20 http://github.com/rasvaan/digibird_client

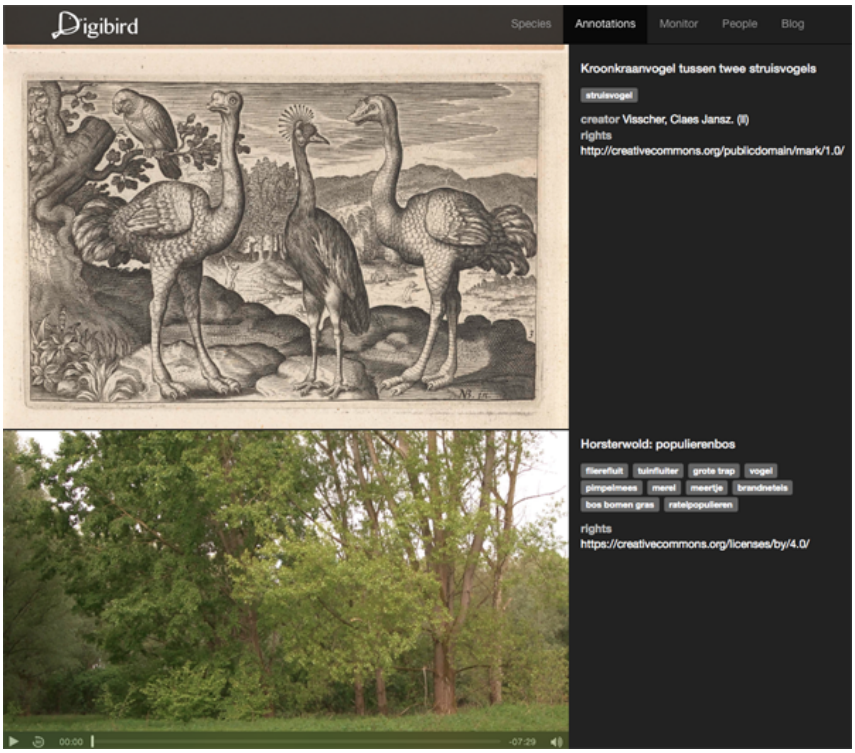

Figure 31: Screenshot of the annotation wall of DigiBird, providing a real-time overview of crowdsourcing results.

from the IOC World Bird List. Results obtained from the participating institutions are shown side by side, providing the user an overview of different types of media available from different sources regarding their search query.

Results are shown the moment they are available, illustrating time differences in retrieving information between the systems. For example, the objects from the Rijksmuseum often take over a minute to be displayed, due to the slow text query. Figure 31 depicts the annotation wall, which shows a real-time overview of crowd contributions. At the moment a contributor adds an annotation in Accurator[21] or tags a video in Waisda?[22], this is immediately displayed on the annotation wall. Additional screenshots of the DigiBird system are shown in Appendix A.4.

As a practical example, the DigiBird API is used by Naturalis Biodiversity Center to embed crowdsourced bird sounds from the Xeno-canto project on bird species information pages of their Dutch Species Register. The benefits are two-directional: Xeno-canto and their contributors have their user-generated content automatically show up in a relevant external portal (causing more plays/hits and opportunities for feedback) and the Dutch Species Register gets fed relevant external content to enrich its species information pages. Building on this

21 http://annotate.accurator.nl
22 http://waisda.beeldengeluid.nl

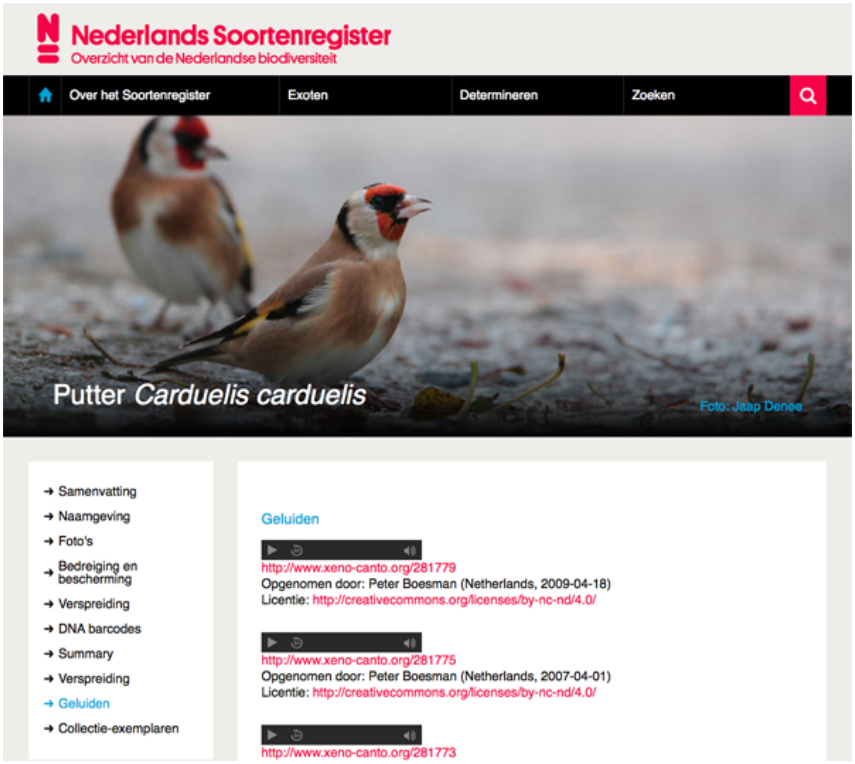

Figure 32: Screenshot of the European Goldfinch page of the Dutch Species Register, with embedded sounds from Xeno-canto.

example, Naturalis Biodiversity Center could also embed prints from the Rijksmuseum with crowdsourced bird identifications. By doing so, a positive feedback loop is created.

6.6 DISCUSSION AND FUTURE WORK

Cultural heritage institutions can benefit greatly from collaborations with other institutions as well as with the public. Considering the current tendency of institutions to publish collection data on the web, online collaborations might seem an easy feat to accomplish. While the DigiBird system manages to integrate four collections extended with crowdsourced data, we had to overcome hurdles regarding data retrieval, linking of objects and data completeness.

Data retrieval and integration are hampered by the diversity of data publishing methods. Within this project, we use APIs as well as public query endpoints. All of them provide access to cultural heritage data, although each one takes a different approach regarding accepted requests and formulated responses. This lack of standardization makes integrating collections from multiple sources a cumbersome task and underlines the importance of documentation. By requesting data from APIs we had to adhere to the limits set by the programmers implementing the API. Uncommon requests had to be

tailor-made or were simply unavailable. The two query endpoints provided a higher level of flexibility, since they allow the formulation of custom requests. However, compared to the APIs, this comes at the cost of overall performance with respect to responsiveness.

The creation of links between objects is required to provide integrated access to objects from different collections. There are two main approaches to achieve this: by creating direct links between objects or by linking objects to a vocabulary shared by multiple collections. The former approach requires the labor-intensive process of creating explicit links between objects every time a new collection is considered. In DigiBird we take the latter approach and we base the integration on a list of bird names, which serves as the "glue" between the collections. Although initiatives such as the Getty Vocabularies [4] improve the availability of concepts to link to, not all topics will have such a clear-cut choice of a suitable vocabulary available.

Crowdsourcing is evolving to be a valuable approach for cultural heritage institutions to collect data. The DigiBird project is a hub for four distinct crowdsourcing projects. Not undertaking these projects in isolation, allows the sharing of resources and provides insights regarding the time needed to collect data. The DigiBird API directly outputs obtained data from the different crowdsourcing projects. As a result, institutions can decide to use data in an early stage of the crowdsourcing process. While this makes it more difficult to ensure a sufficient level of quality, it contributes to a sense of progress and gratification of the crowd contributing to the projects. This DigiBird API is currently used by the Dutch Species Register, embedding content from other sources on the website.

In the future, DigiBird can contribute to the challenge of addressing the promotional effort required to successfully run a crowdsourcing campaign. Based on available data, an assessment can be made of the completeness of metadata. If the metadata of an object is deemed incomplete, a tailored crowdsourcing task could be formulated. Sharing these tasks through an API would allow institutions to embed crowdsourcing tasks on their websites in a similar fashion as collection objects. Thereby moving beyond collaborations on the level of collection items towards engaging audiences in an integrated crowdsourcing approach.

7

CONCLUSIONS

In this thesis, we presented a reusable method to contextualize and enrich cultural heritage collections, to support explorative search and collection integration. The reusable method consists of five steps, each step generating input for subsequent steps. We applied the method to the collection of the Rijksmuseum Amsterdam. We started by setting a baseline: analyzing the available collection data. Data modeling challenges encountered during this analysis served as input for a comparison of modeling approaches. To add more contextual information to objects, we introduced a nichesourcing method and used it to collect enrichments. We continued by investigating how enrichments from different structured vocabularies influence the ability to explore online cultural heritage collections. Additionally, we investigated how continuous enrichment of collections can be used to integrate data from multiple institutions. In this chapter, we revisit each of these topics by answering the research questions posed in Chapter 1. Also, we discuss the implications of our work and propose directions for future work.

7.1 RESEARCH QUESTIONS REVISITED

We introduced four research questions in Chapter 1, regarding contextualization, enrichment, exploration and integration. The origin of the first research question is an analysis of the Linked Data published by the Rijksmuseum. A conversion maps data from the collection management system to two standardized data models: the Dublin Core metadata scheme and the Europeana Data Model. During this conversion information is lost, due to the limited capability of the models to capture contextual information. Following these findings, we investigated the impact of modeling approaches on capturing contextual information, in the first research question:

> *1. How do different modeling approaches influence the contextualization of cultural heritage collections published online?*

When institutions consider publishing data structured using an ontology, they need to assess whether the ontology can adequately capture information about their objects. Based on the analysis of the Rijksmuseum Linked Data as well as related work, we identified six modeling challenges. In Chapter 3, each of these challenges is illustrated with modeling approaches encountered in two commonly used cul-

tural heritage ontologies: the Europeana Data Model and the CIDOC Conceptual Reference Model.

The modeling approaches taken by ontologies do influence what information can be captured. From the modeling challenges and approaches, we derived six requirements for cultural heritage ontologies. These requirements regard specialization, object- and event-centric approaches, temporality, representations, views and subject matter. An ontology that addresses each requirement will become more complex. By considering use cases, institutions should decide which constructs of ontologies are needed, thereby striking the right balance between expressivity and complexity.

One of the use cases to consider is the description of subject matter of objects. The required knowledge to adequately describe subject matter is sometimes out of the scope of the institutional expertise. The second research question, therefore, regards how external experts can be involved in the process of contributing contextual information:

> *2. What method for engaging niche communities to enrich cultural heritage objects can result in high-quality annotations?*

To investigate enrichment strategies, we introduced the Accurator nichesourcing methodology in Chapter 4. Accurator involves niche communities in the annotation process of cultural heritage objects. We define niche communities as groups of intrinsically motivated lay-experts, knowledgeable about highly specialized topics. The Accurator methodology enables institutions to tap into this knowledge pool. The methodology consists of four stages: an orientation, implementation, execution and evaluation stage. In total, we ran three case studies to validate the methodology.

All of the case studies resulted in high-quality annotations. Responses to a questionnaire indicated that the contributors enjoyed working together and found that the Accurator annotation tool supported them well in adding information to objects. Therefore, we conclude that Accurator is a methodology that addresses, supports and motivates niche communities, allowing institutions to collect high-quality annotations. We continued by investigating how such enrichments influence the ability to explore collections in the third research question:

> *3. How do enrichments from various structured vocabularies influence the diversity of search results?*

Explorative search can be used to reach relevant objects in a cultural heritage collection, while the exact contents of the collection are unknown to the user. In Chapter 5, we investigated whether enrichments from structured vocabularies have an impact on the diversity of search results generated by semantic search algorithms. From the logs of the Rijksmuseum website, we obtained a set of search queries. We

used a semantic search algorithm, which retrieves objects connected with the search query through a path in the data. Next, the objects are clustered based upon the paths used to retrieve them. Thereby, the search algorithm leverages the connections between objects and concepts in structured vocabularies. To answer the research question, we set up an experiment in which we, turn by turn, loaded enrichments from five different structured vocabularies, aligned with the Rijksmuseum collection data. We used the length of paths and the number of clusters as indicators of search result diversity.

Based on the increased path lengths and number of clusters, we conclude that the enrichments result in an increased diversity of search results. Also, the enrichments allowed the search algorithm to obtain more search results. Taking these findings into account, we hypothesized which characteristics of the datasets increase the diversity of search results. First, the number of internal connections between concepts of a vocabulary seems to influence the ability to form paths. Second, the number of connections between objects and distinct concepts from structured vocabularies seems to increase the diversity of search results. The nichesourcing methodology described in Chapter 4 contributes to the number of added concepts and therefore contributes to the possibility for users to explore cultural heritage collections. In the next research question we investigate the challenges related to integrating multiple collections enriched using crowdsourcing:

> *4. How to address the issue of continuously evolving data in the process of integrating cultural heritage datasets from various institutions?*

Different online cultural heritage collections often contain complementary objects, which makes integration a worthwhile effort. In Chapter 6, we integrate four nature-related collections, containing objects of three different modalities: sounds, images and videos. All four collections are enriched by crowdsourcing initiatives, either by allowing contributors to add objects or by asking contributors to enrich the data about objects. These initiatives generate a continuous stream of added information, which in turn is beneficial for collection exploration and integration. To leverage the added information, we had to find a way to integrate constantly changing datasets.

To answer the research question, we build a system that integrates collections and is able to cope with data that constantly changes. The collections and crowdsourced annotations were stored in different systems, each one of them taking a different approach regarding accepted requests and formulated responses. This lack of standardization required us to tailor make retrieval and harmonization methods. The two sources that use Semantic Web technologies made data retrieval straightforward, although the response time of some queries negatively affected the user experience.

We created a user interface, that provides an integrated view on artworks, images, sounds and videos about a specified type of species. Additionally, the progress of the crowdsourcing initiatives can be monitored. The final DigiBird system brought together all the results presented in this thesis, integrating enriched Rijksmuseum Linked Data with other sources and improving the search experience for users.

7.2 DISCUSSION

In this section, we discuss the generalizability of our approach, in addition to five overarching themes. The first theme is the data-driven nature of this research, which deviates from the often object focused research in the cultural heritage domain. Data-driven research is enabled by the growing tendency of institutions to publish data online about their collections. The second theme regards the publication of cultural heritage data and how institutions can make the most of this upcoming practice. Nichesourcing can be used to enrich collections, but also serves as a way to engage with communities. The engagement of institutions with the public is the third theme we discuss. The availability of collections and structured vocabularies changes the data landscape an institution operates in. The fourth theme regards how institutions can cope with these new sources of contextual information. The Accurator and DigiBird systems use collection data and structured vocabularies. The last theme regards how systems originating from research initiatives can be turned into software part of production environments. Before we discuss these themes, we first consider the generalizability of the five-step approach.

GENERALIZABILITY

All studies presented in this thesis are conducted in the cultural heritage domain and most of them in the context of the collection of the Rijksmuseum Amsterdam. This focus prohibits us from making definitive claims about the applicability of our findings outside this domain, although we expect that many steps of the method can be utilized in cases from other domains as well. The applicability and usefulness of steps will depend on the characteristics of the case in question. Relevant characteristics concern the dataset, the context of the dataset and the community. We will briefly discuss each of these below.

At the foundation of the five-step method lies the availability of a dataset with metadata about objects. The metadata can describe real-world objects or born-digital objects, although digital representations of the objects should be available to be annotated. The datasets are by no means limited to cultural heritage collections. Examples of other usable datasets are a product catalog of an online shop or

sound clips from a radio broadcaster. In this research, we dealt with a single modality of representations: images. We expect that the Accurator nichesourcing methodology can also be used for other media modalities, even though the Accurator annotation tool will have to be adapted. A broadcaster might need to, for example, extend the tool with functionality to support annotating sound or video clips.

The availability of domain ontologies, structured vocabularies and related collections influence the generalizability of the contextualization of datasets. Domain ontologies have to be available to compare different approaches to structure contextual information about the objects. If no such standardized ontology is available, this will hamper the possibility to integrate the dataset with related data sources. Furthermore, the degree in which structured vocabularies cover the domain influences the accuracy of enrichments. As discussed in this thesis, there is a number of high-quality vocabularies available in the cultural heritage domain, which is not the case for every domain. For the domains that lack appropriate vocabularies, more generic external datasets could be used, such as WikiData [79]. Likewise, related datasets have to be available before a dataset can be contextualized, by integrating it with related sources.

An active community interested in the topic at hand should exist for the five-step methodology to be effective. This is especially true for the annotation step, which relies on the voluntary contribution of niche groups. Not every case will have a community cut out for it, although we recommend requesters to be inventive in the selection of communities. A broadcaster looking to describe video clips might, for example, involve locals knowledgeable about the situation. The goal of the methodology is to improve access to online collections. Improvements correlate with use cases of the data, that again hinge on the availability of users.

DATA-DRIVEN RESEARCH IN THE CULTURAL HERITAGE DOMAIN

The research reported upon in this thesis is data-driven, every analysis is based on real-world data. We started by analyzing the data published by the Rijksmuseum, which accumulated the source of this dataset over the past decades, by entering information about collection objects in the collection management system. New insights and procedures changed how collection objects have been described over the years, resulting in descriptions of varying quality (Chapter 2). By analyzing the collection, we were able to identify information gaps and actively address them with methods such as nichesourcing. In Chapter 5, we investigated whether enrichments impact the diversity of obtained search results. Again, this analysis is based on real-world data, by using the query logs of the Rijksmuseum website.

Employees of cultural heritage institutions often work on the level of single objects. Each of these object descriptions individually might

seem correct, but analysis at the level of sub-collections might reveal inconsistencies. The insights offered by statistics gained at a collection and structured vocabulary level are useful for guiding improvements of collection registration procedures. In addition, the uptake of standardized ways of publishing data allows for analyzing data at the level of many datasets at once [65]. At the moment more cultural heritage collections become available as Linked Data, analysis of multiple collections will provide new insights in collection registration practices. Aligning these practices within the domain will ease collection integration efforts such as DigiBird.

PUBLICATION OF CULTURAL HERITAGE DATA

Disseminating high-quality information about objects is embedded in the mission of cultural heritage institutions [78]. As shown in Chapter 2, publishing data increases the visibility of collections. To benefit from this, it should be clear who is the source of the data. This is currently obfuscated by the use of third parties to host data and provide persistent identifier services. Data is often published on platforms outside the institutions' own infrastructure, such as aggregators. The increased distance between the institution and its data lowers the sense of ownership and responsibility, while making it hard to identify the source for users. One of the definitions for *authority* in the Oxford Dictionary reads: "A book or other source able to supply reliable information or evidence". To reclaim authority, institutions should host their own data and provide their own persistent identifier services.

Linked Open Data can be used by anyone, for anything. To support as many different applications as possible, an institution has to strike the right balance between simplicity and expressivity of published data. Usage of standardized data models can provide guidance, although, as seen in Chapter 3, data models impact expressivity. An institution can choose to support multiple data models, but even better, it could investigate how data is used and adapt based on the requirements of users. Alterations can be made by extending the source or improving the conversion of the source. The Rijksmuseum engaged with users by participating in research projects and aggregation initiatives, which helped shape the data (Chapter 2). We continue by discussing how nichesourcing is used to extend source information and to engage with new audiences.

ENGAGEMENT OF THE PUBLIC

Next to being a method for outsourcing tasks, crowdsourcing can be used by cultural heritage institutions to engage with the public [64]. In Chapter 4, we helped to intensify this engagement with a nichesourcing method, by inviting groups of experts to come visit the institution and contribute their knowledge. Niche communities are

social entities and this resulted in lively discussions during the events. These discussions were useful for the further development of collection registration methods regarding the topic at hand. However, we did not yet manage to integrate this social aspect into our online annotation tool. A step towards this goal would be to build in active collaboration options and the possibility to discuss tasks. Thereby, potentially resulting in a permanent positive feedback loop of engagement and enrichment.

An often asked question by contributors is what happened with their contributions. Knowing that their work matters helps to keep intrinsically motivated contributors involved. In the Accurator annotation tool described in Section 4.4, we show the direct impact of annotations on search functionality. Additionally, the DigiBird system showed the impact of annotations on collection integration (Chapter 6). Still, it would be even better if the results of crowdsourcing tasks would be directly visible on the websites of institutions. Therefore, we continue by discussing how institutions could better cope with such new sources of information.

COPING WITH NEW SOURCES OF CONTEXTUAL INFORMATION

Cultural heritage institutions that want to publish rich contextualized data online, will have to adapt to a new data landscape. Institutions used to rely on internal processes to generate information about their collection, which allowed strict control over quality. With the increase of available digitized content, manual processes to describe content do not suffice anymore. The rise of new data generation methods (e.g. machine learning, crowdsourcing) and contextual data sources (e.g. structured vocabularies) provide the means to address this issue. However, to fully embrace these methods and sources, challenges regarding data quality need to be addressed.

Two main aspects impact the quality of contextual information in the cultural heritage domain: the method of relating collection data to contextual data sources and the quality of contextual data source themselves. Institutions have to carefully consider which contextual sources of information to use, as sources often reside outside the sphere of influence of the institution. This can be problematic because it makes it more difficult to add new concepts, correct existing concepts or change the hierarchy of concepts. To illustrate, most contextual sources are limited in scope and this scope might not be aligned with the needs of an institution. For example, the Iconclass vocabulary, used by the Rijksmuseum to describe subject matter, focusses on Western art, while the museum also has many Asian objects. To cover the complete spectrum of contextual information of a collection, institutions have to keep evaluating the list of structured vocabularies in use and consider new sources when they become available.

This thesis proposes different methods to link collection objects to structured vocabularies. First, we analyzed links already made by employees of the Rijksmuseum in Chapter 2. The number of links was low: only a fraction of contextual information stems from external vocabularies. Second, we automatically generated links based on a simple string matching algorithm, to investigate the added value of an increased number of links in the context of search (Chapter 5). This method relied on existing textual information and does not cope well with ambiguous terms, inevitably introducing incorrect links. Third, to overcome quality and quantity issues, we introduced a nichesourcing method in Chapter 4, inviting experts to link collection objects to structured vocabularies.

The automatically generated links and the links provided by the niche experts have not yet been added to the collection management system of the Rijksmuseum. To better facilitate this ingestion process, we need to provide institutions with tooling to assess the quality of the generated links. Knoblock et al. [50] created a tool to assess every proposed link made. Ceolin, Nottamkandath, and Fokkink [10] formulated automated methods to evaluate the quality of crowdsourced annotations, based on trust assessments. In the future, the Accurator annotation tool could be extended with such automated methods. As a result, institutions could make better-informed decisions about whether to ingest the crowdsourced data, with less effort.

FROM PROOF-OF-CONCEPTS TO PRODUCTION-READY SYSTEMS

Both cultural heritage institutions, as well as research institutes, can take steps to ease the translation of research results into useful applications for the cultural heritage domain. This thesis describes two systems, that would be of potential use for many institutions. While we published the source code of both, uptake of the systems has been limited. Many institutions showed interest in using the Accurator annotation tool, but the relatively difficult deployment proved to be a significant barrier. Institutions have to realize that codebases originating from research projects are often proofs-of-concept, which are of limited use in production environments.

The Accurator annotation system had to be developed beyond a proof-of-concept, since we ask contributors to work with it for longer periods of time and rely on their intrinsic motivation. If the system would not have been stable and usable, this would hold back experts in contributing information. The DigiBird system is a proof-of-concept for on the fly collection integration, relying on other systems for data. The importance of standardization is illustrated by the effort needed to adapt pipelines to each individual API specification. An additional problem was the instability of API specifications, regularly resulting in broken integration pipelines. To support similar

collection integration initiatives in the future, the API and Linked Data eco-systems will need to mature and stabilize.

To foster technological advancements and grow proofs-of-concept into full-blown production-ready systems, significant development steps need to be taken. To ease this process, larger institutions can consider "lab environments"[1]. These environments showcase new experimental technology, without having to conform to the expected stability of production systems. If the technology is promising enough, production-ready implementations can be developed by parties outside academia. If academics would consider this trajectory in an early stage of the proof-of-concept development, more results could be translated into production-ready systems.

1 For example, Europeana, the National Library of the Netherlands and the Netherlands Institute of Sound and Vision Institute have lab environments: `http://europeanalabs.eu`, `http://lab.kb.nl` and `http://labs.beeldengeluid.nl`

A

SCREENSHOTS

In this appendix we show screenshots of the Accurator annotation tool, the DigiBird system and related systems.

A.1 BIRDS ON ART ACCURATOR ANNOTATION TOOL

Screenshots of the Accurator annotation tool instantiated for the "Birds on art" case study of Section 4.5.1.

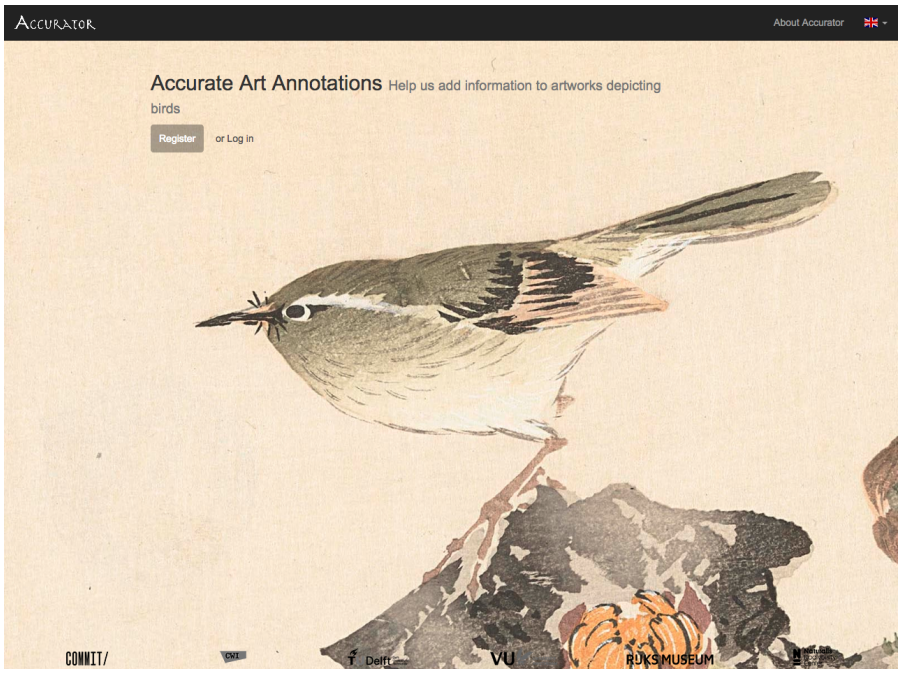

Figure 33: Screenshot of the home page of the "Birds on art" domain.

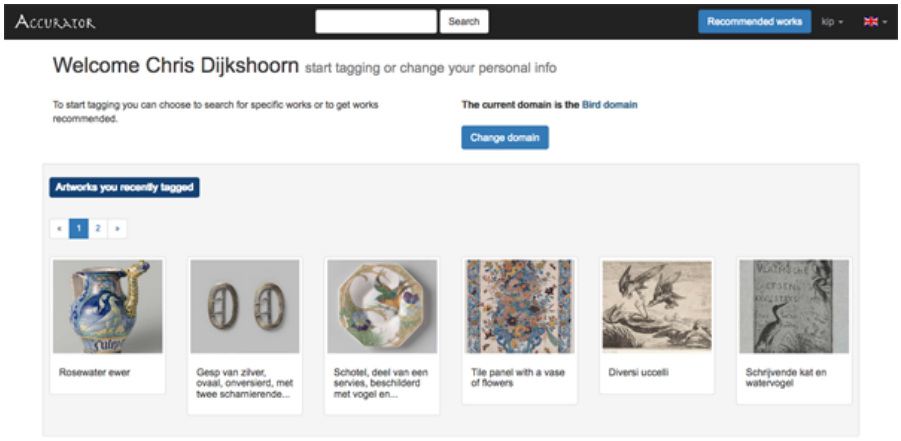

Figure 34: Screenshot of the profile page.

A.2 BIBLE PRINTS ACCURATOR ANNOTATION TOOL

Screenshots of the Accurator annotation tool instantiated for the "Bible prints" case study of Section 4.5.2.

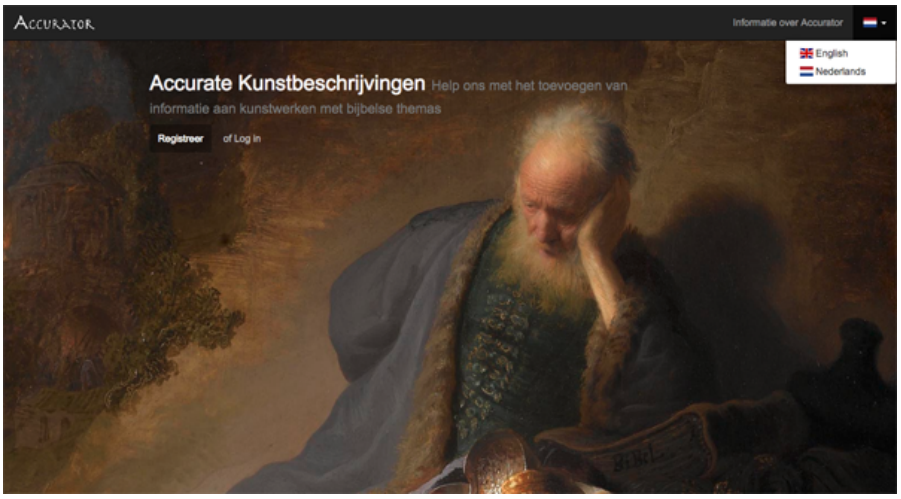

Figure 35: Screenshot of the Dutch version of the "Bible prints" domain.

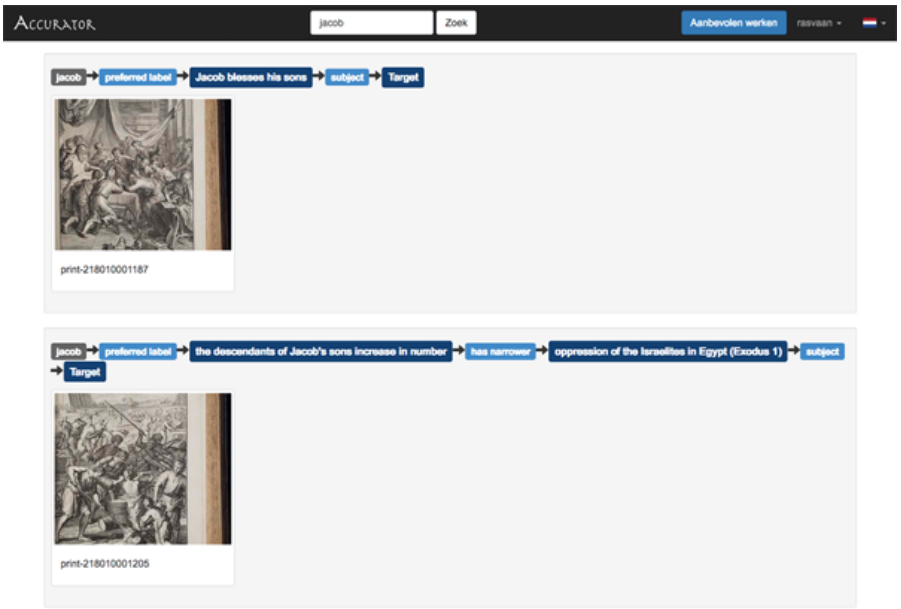

Figure 36: Screenshot of the search results for the query "Jacob".

A.3 FASHION IMAGES ACCURATOR ANNOTATION TOOL

Screenshots of the Accurator annotation tool instantiated for the "Fashion images" case study of Section 4.5.3.

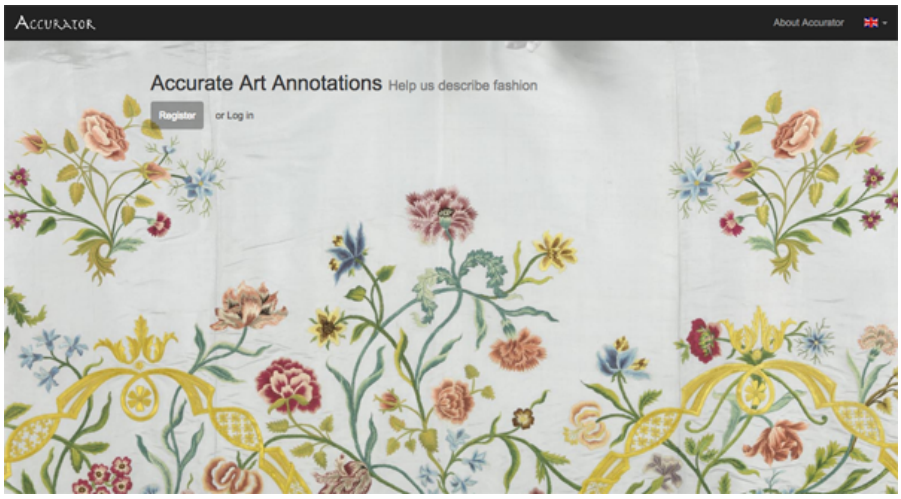

Figure 37: Screenshot of the home page of the "Fashion images" domain.

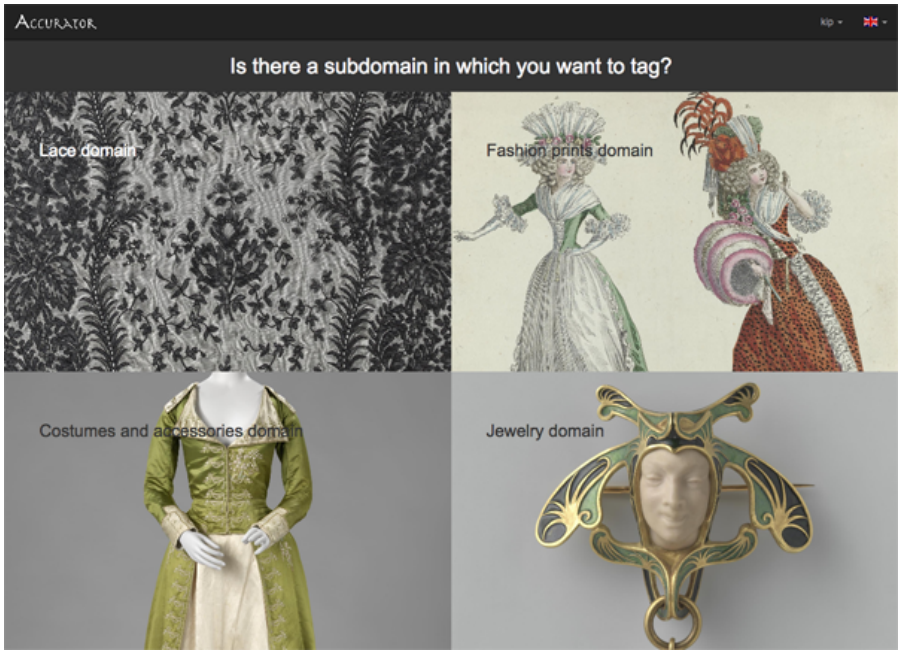

Figure 38: Screenshot of subdomain selection.

A.4 SCREENSHOTS DIGIBIRD SYSTEM

In this appendix we show screenshots of the DigiBird collection integration system, as described in Chapter 6.

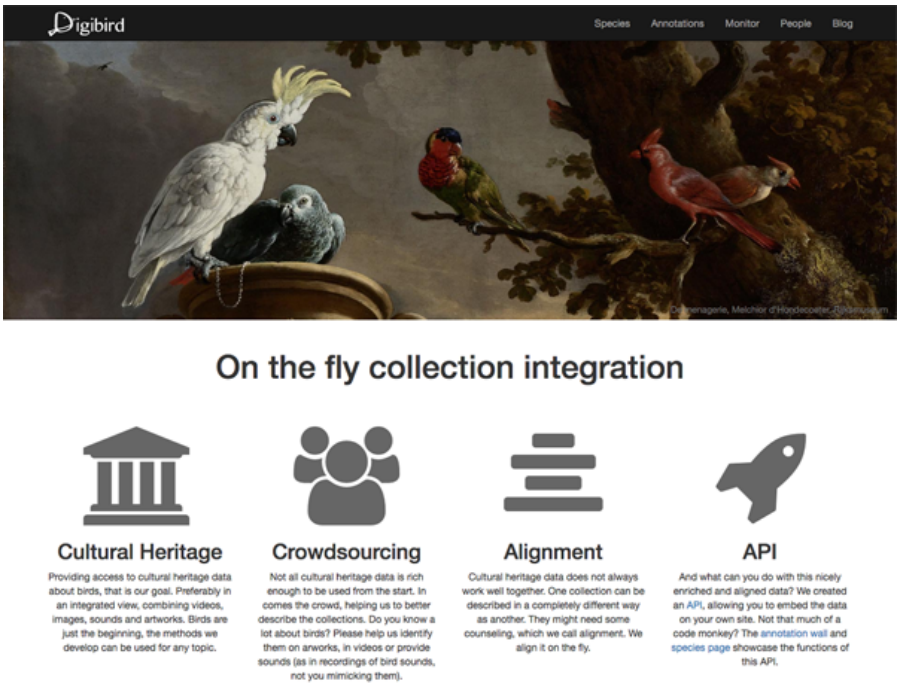

Figure 39: Screenshot of the DigiBird homepage.

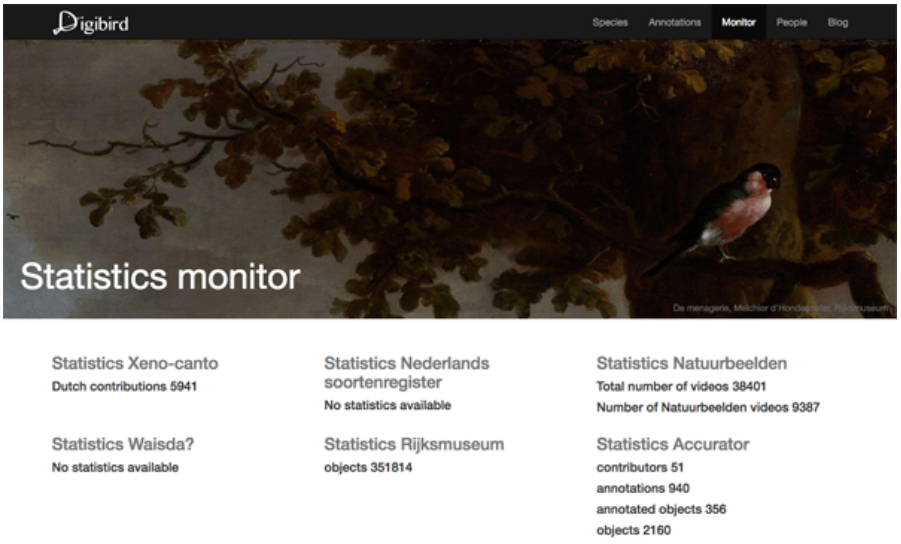

Figure 40: Screenshot of the statistics monitor.

A.5 SCREENSHOTS OF RELATED SYSTEMS

In this appendix we show screenshots of systems related to the Digi-Bird system, as mentioned in Section 6.2.

Figure 41: Screenshot of Natuurbeelden homepage.

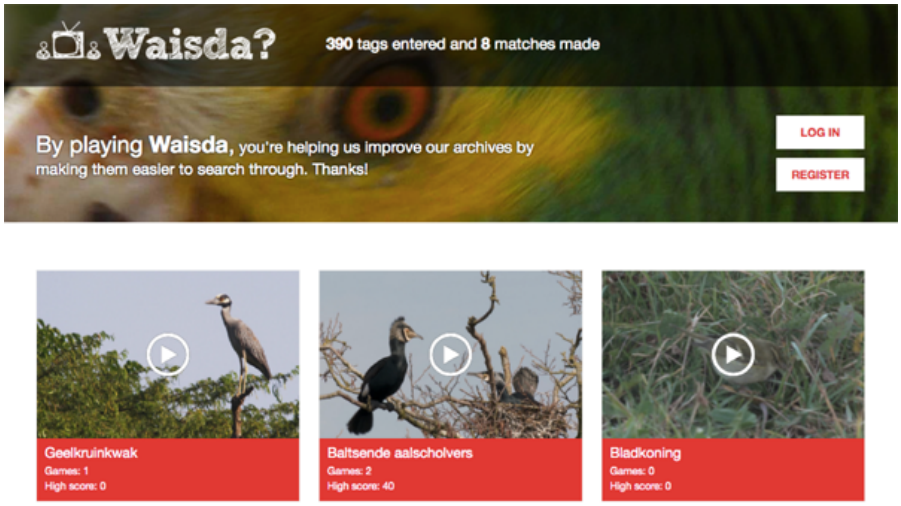

Figure 42: Screenshot of Waisda? system.

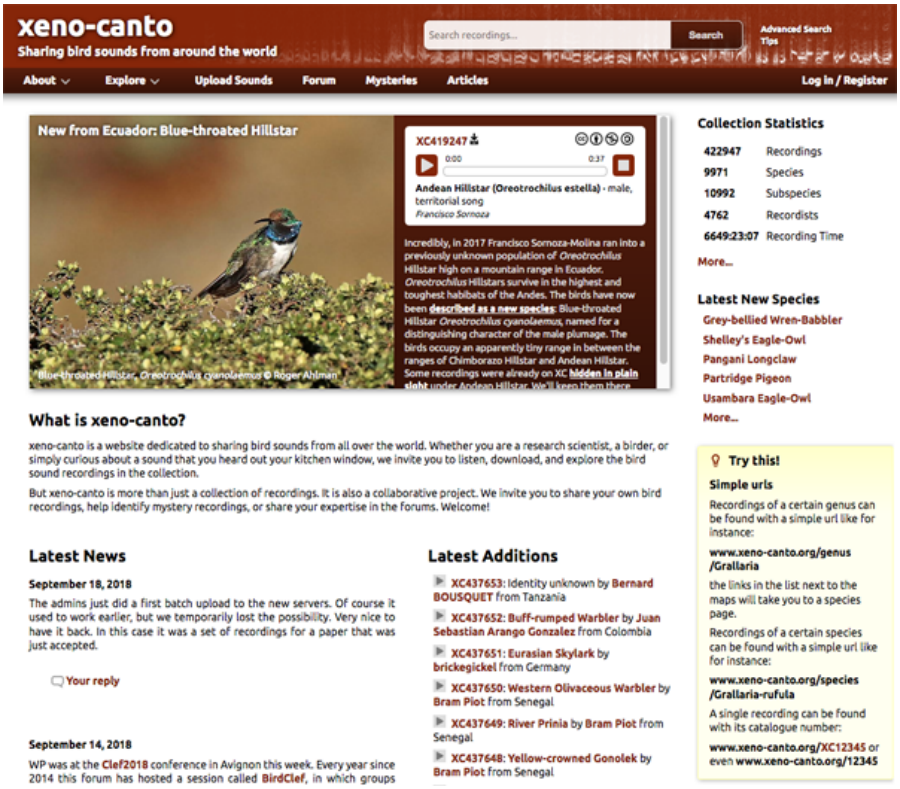

Figure 43: Screenshot of the Xeno-canto homepage.

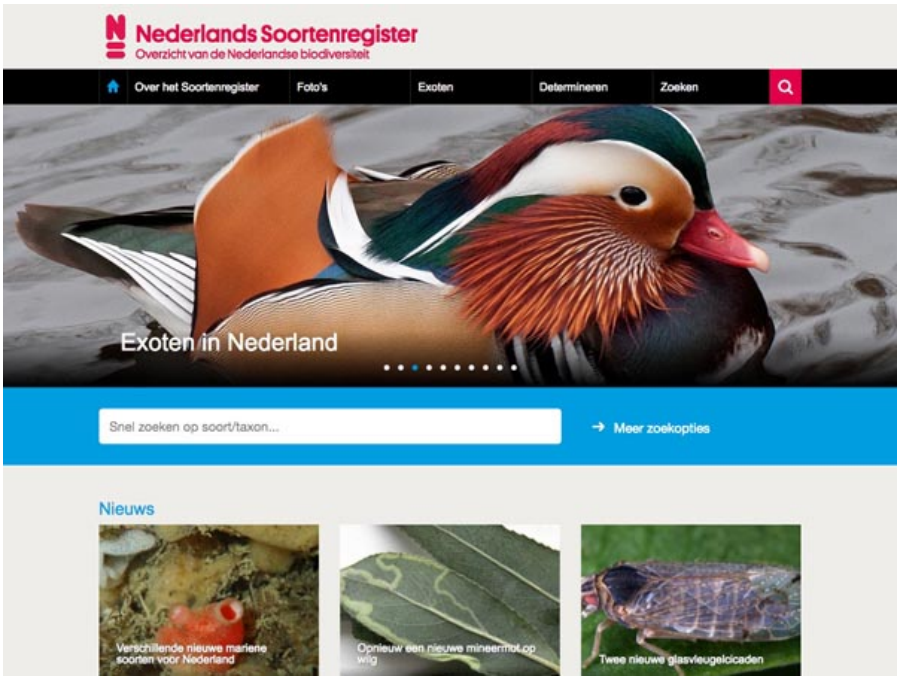

Figure 44: Screenshot of the Nederlandse Soortenregister homepage.

BIBLIOGRAPHY

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. "Diversifying search results." In: *Proceedings of the 2nd acm international conference on web search and data mining*. (Barcelona, Spain). Ed. by Ricardo Baeza-Yates, Paolo Boldi, Berthier Ribeiro-Neto, and B. Barla Cambazoglu. WSDM '09. ACM, 2009, pp. 5–14. DOI: 10.1145/1498759.1498766.

[2] Luis von Ahn and Laura Dabbish. "Labeling images with a computer game." In: *Proceedings of the sigchi conference on human factors in computing systems*. (Vienna, Austria). CHI '04. ACM, 2004, pp. 319–326. DOI: 10.1145/985692.985733.

[3] Chiel van den Akker, Susan Legêne, Marieke van Erp, Lora Aroyo, Roxane Segers, Lourens van der Meij, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. "Digital hermeneutics: Agora and the online understanding of cultural heritage." In: *Proceedings of the 3rd international web science conference*. (Koblenz, Germany). WebSci '11. ACM, 2011, pp. 1–7. DOI: 10.1145/2527031.2527039.

[4] Murtha Baca and Melissa Gill. "Encoding multilingual knowledge systems in the digital age: the getty vocabularies." In: *Knowledge organization* 42.4 (2015), pp. 232–243. ISSN: 09437444.

[5] Thomas Baker, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber, and Ed Summers. "Key choices in the design of Simple Knowledge Organization System (SKOS)." In: *Web semantics: science, services and agents on the world wide web* 20 (2013), pp. 35–49. ISSN: 1570-8268. DOI: 10.1016/j.websem.2013.05.001.

[6] Tim Berners-Lee, James Hendler, and Ora Lassila. "The Semantic Web." In: *Scientific american* 284.5 (2001), pp. 28–37.

[7] Andrew Bevan, Daniel Pett, Chiara Bonacchi, Adi Keinan Schoonbaert, Daniel Lombraña González, Rachael Sparks, Jennifer Wexler, and Neil Wilkin. "Citizen archaeologists. online collaborative research about the human past." In: *Human computation journal* 1.2 (2014), pp. 185–199. ISSN: 2330-8001. DOI: 10.15346/hc.v1i2.9.

[8] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked Data: the story so far." In: *Semantic services, interoperability and web applications: emerging concepts*. Ed. by Amit Sheth. IGI Global, 2011. Chap. 8, pp. 205–227. DOI: 10.4018/978-1-60960-593-3.ch008.

[9] Maarten Brinkerink, Chris Dijkshoorn, Henrike Hövelmann, Sander Pieterse, and Maarten Heerlien. *Bridging the natural divide: crowd-curation of cultural expressions inspired by nature*. Panel presentation MCN2014, Open Data/Participation Track. 2014.

[10] Davide Ceolin, Archana Nottamkandath, and Wan Fokkink. "Automated evaluation of annotators for museum collections using subjective logic." In: *Proceedings of the 6th ifip trust management conference*. (Surat, India). IFIPTM '12. Springer. May 2012, pp. 232–239. DOI: 10.1007/978-3-642-29852-3_18.

[11] Jon Chamberlain. "Groupsourcing: distributed problem solving using social networks." In: *Proceedings of the 2nd aaai conference on human computation and crowdsourcing*. (Pittsburgh, PA, USA). HCOMP '14. The AAAI Press, Nov. 2014, pp. 22–29.

[12] Susan Chun, Rich Cherry, Doug Hiwiller, Jennifer Trant, and Bruce Wyman. "Steve.museum: an ongoing experiment in social tagging, folksonomy, and museums." In: *Proceedings of the museums and the web conference*. Ed. by Jennifer Trant and David Bearman. 2006. URL: http://www.archimuse.com/mw2006/papers/wyman/wyman.html.

[13] Robina Clayphan, Valentine Charles, and Antoine Isaac. *Europeana Data Model – mapping guidelines*. 2.3. Europeana. Nov. 2016. URL: http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.3_112016.pdf.

[14] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. "SuggestBot: using intelligent task routing to help people find work in Wikipedia." In: *Proceedings of the 12th international conference on intelligent user interfaces*. (Honolulu, Hawaii, USA). IUI '07. ACM Press, Jan. 2007, pp. 32–41. DOI: 10.1145/1216295.1216309.

[15] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. *Definition of the CIDOC conceptual reference model*. 5.0.4. CIDOC CRM Special Interest Group. Nov. 2011.

[16] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. "Pick-a-crowd: tell me what you like, and I'll tell you what to do." In: *Proceedings of the 22nd international conference on world wide web*. (Rio de Janeiro, Brazil). WWW '13. ACM, 2013, pp. 367–374. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488421.

[18] Chris Dijkshoorn, Mieke H. R. Leyssen, Archana Nottamkandath, Jasper Oosterman, Myriam Traub, Lora Aroyo, Alessandro Bozzon, Wan Fokkink, Geert-Jan Houben, Henrike Hovelmann, Lizzy Jongma, Jacco van Ossenbruggen, Guus Schreiber,

and Jan Wielemaker. "Personalized nichesourcing: acquisition of qualitative annotations from niche communities." In: *Workshop proceedings of the 21st conference on user modeling, adaptation, and personalization*. Ed. by Shlomo Berkovsky, Eelco Herder, Pasquale Lops, and Olga C. Santos. Vol. 997. CEUR Workshop Proceedings. CEUR-WS.org, 2013.

[19] Chris Dijkshoorn, Lora Aroyo, Guus Schreiber, Jan Wielemaker, and Lizzy Jongma. "Using Linked Data to diversify search results: a case study in cultural heritage." In: *Proceedings of the 19th international conference on knowledge engineering and knowledge management*. (Linköping, Sweden). Ed. by Krzysztof Janowicz, Stefan Schlobach, Patrick Lambrix, and Eero Hyvönen. EKAW '14. Springer International Publishing, Nov. 2014, pp. 109–120.

[20] Chris Dijkshoorn, Cristina-Iulia Bucur, Maarten Brinkerink, Sander Pieterse, and Lora Aroyo. "DigiBird: on the fly collection integration supported by the crowd." In: *Proceedings of the museums and the web conference*. Apr. 2017.

[21] Chris Dijkshoorn, Lora Aroyo, Jacco Van Ossenbruggen, and Guus Schreiber. "Modeling cultural heritage data for online publication." In: *Applied ontology* 13.4 (2018), pp. 255–271.

[22] Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco Van Ossenbruggen, Guus Schreiber, Wesley ter Weele, and Jan Wielemaker. "The Rijksmuseum collection as Linked Data." In: *Semantic web journal* 9.2 (2018), pp. 221–230.

[23] Chris Dijkshoorn, Victor de Boer, Lora Aroyo, and Guus Schreiber. "Accurator: nichesourcing for cultural heritage." In: *Human computation journal* (in press).

[24] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. "Crowdsourcing systems on the world-wide web." In: *Communications of the acm* 54.4 (Apr. 2011), pp. 86–96. ISSN: 0001-0782. DOI: 10.1145/1924421.1924442.

[25] Martin Doerr. "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata." In: *Ai magazine* 24.3 (2003), pp. 75–92.

[26] Martin Doerr and Nicholas Crofts. "Electronic esperanto: the role of the object oriented CIDOC reference model." In: *Proceedings of the cultural heritage informatics conference*. (Washington DC). ICHIM '99. Sept. 1999, pp. 22–26.

[27] Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert van de Sompel. "The Europeana Data Model (EDM)." In: *World library and information congress: 76th ifla general conference and assembly*. 2010, pp. 10–15.

[28] Mauro Dragoni, Elena Cabrio, Sara Tonelli, and Serena Villata. "Enriching a small artwork collection through semantic linking." In: *Proceedings of the 13th extended semantic web conference*. (Heraklion, Crete, Greece). Ed. by Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange. ESWC '16. Springer International Publishing, May 2016, pp. 724–740. ISBN: 978-3-319-34129-3. DOI: 10.1007/978-3-319-34129-3_44.

[29] Alastair Dunning and Ingeborg Verspille. *The current Europeana dataset*. Mar. 2017. URL: http://research.europeana.eu/about-our-data/the-current-europeana-dataset.

[30] Andrew Ellis, Dan Gluckman, Adrian Cooper, and Greg Andrew. "Your Paintings: a nation's oil paintings go online, tagged by the public." In: *Proceedings of the museums and the web conference*. 2012. URL: http://www.museumsandtheweb.com/mw2012/papers/your_paintings_a_nation_s_oil_paintings_go_onl.

[31] Joost Geurts, Stefano Bocconi, Jacco van Ossenbruggen, and Lynda Hardman. "Towards ontology-driven discourse: from semantic graphs to multimedia presentations." In: *Proceedings of the 2nd international semantic web conference*. Ed. by Dieter Fensel, Katia P. Sycara, and John Mylopoulos. ISWC '03. Springer, 2003, pp. 597–612. DOI: 10.1007/978-3-540-39718-2_38.

[32] Riste Gligorov, Michiel Hildebrand, Jacco van Ossenbruggen, Guus Schreiber, and Lora Aroyo. "On the role of user-generated metadata in audio visual collections." In: *Proceedings of the 6th international conference on knowledge capture*. (Banff, Alberta, Canada). K-CAP '11. ACM Press, May 2011, pp. 145–152. DOI: 10.1145/1999676.1999702.

[33] Riste Gligorov, Michiel Hildebrand, Jacco Ossenbruggen, Lora Aroyo, and Guus Schreiber. "An evaluation of labelling-game data for video retrieval." In: *Proceedings of the 35th european conference on ir research*. (Moscow, Russia). Ed. by Pavel Serdyukov, Pavel Braslavski, SergeiO. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz. ECIR '13. Springer, Mar. 2013, pp. 50–61. DOI: 10.1007/978-3-642-36973-5_5.

[34] Peter Gorgels. "Rijksstudio: make your own masterpiece!" In: *Proceedings of the museums and the web conference*. Ed. by Nancy Proctor and Rich Cherry. Jan. 2013. URL: http://mw2013.museumsandtheweb.com/paper/rijksstudio-make-your-own-masterpiece/.

[35] Shinsuke Goto, Toru Ishida, and Donghui Lin. "Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes." In: *Proceedings of the 4th aaai confer-*

ence on human computation and crowdsourcing. (Austin, TX, USA). HCOMP '16. The AAAI Press, Nov. 2016, pp. 52–58.

[36] Andrew Greg. "Your Paintings: public access and public tagging." In: *Journal of the scottish society for art history* 16 (2012), pp. 48–52.

[37] Gunnar Aastrand Grimnes, Peter Edwards, and Alun D. Preece. "Instance based clustering of semantic web resources." In: *Proceedings of the 5th european semantic web conference*. (Tenerife, Canary Islands, Spain). Ed. by Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis. Vol. 5021. ESWC '08. Springer Berlin Heidelberg, 2008, pp. 303–317. DOI: 10.1007/978-3-540-68234-9_24.

[38] Nicola Guarino, Daniel Oberle, and Steffen Staab. "What is an ontology?" In: *Handbook on ontologies*. Ed. by Steffen Staab and Rudi Studer. Springer, 2009, pp. 1–17. ISBN: 978-3-540-92673-3. DOI: 10.1007/978-3-540-92673-3_0.

[39] Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman, and Geertje Jacobs. "Supporting subject matter annotation using heterogeneous thesauri: a user study in web data reuse." In: *International journal of human-computer studies* 67.10 (2009), pp. 887–902. DOI: 10.1016/j.ijhcs.2009.07.008.

[40] Laura Hollink, Guus Schreiber, and Bob Wielinga. "Patterns of semantic relations to improve image content search." In: *Web semantics: science services and agents on the world wide web* 5.3 (2007), pp. 195–203. ISSN: 15708268. DOI: 10.1016/j.websem.2007.05.002.

[41] Laura Hollink, Guus Schreiber, Bob Wielinga, and Marcel Worring. "Classification of user image descriptions." In: *International journal of human-computer studies* 61.5 (Nov. 2004), pp. 601–626. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2004.03.002.

[42] Seth van Hooland and Ruben Verborgh. *Linked Data for libraries, archives and museums*. Facet Publishing, 2014.

[43] Eero Hyvönen, Eetu Mäkelä, Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila, and Suvi Kettula. "Museumfinland – finnish museums on the semantic web." In: *Web semantics: science, services and agents on the world wide web* 3.2-3 (2005), pp. 224–241. DOI: http://dx.doi.org/10.1016/j.websem.2005.05.008.

[44] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. "Crowdtruth: machine-human computation framework for harnessing disagreement in gathering annotated data." In: *Proceedings of the 13th international semantic web conference*. (Riva del Garda, Italy). Ed. by Peter Mika, Tania

Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble. ISWC '14. Springer, 2014, pp. 486–504. DOI: 10.1007/978-3-319-11915-1_31.

[45] Antoine Isaac. *Europeana data model primer*. Europeana. June 2013. URL: http://pro.europeana.eu/page/edm-documentation.

[46] Antoine Isaac. *Definition of the Europeana data model*. 5.2.6. Europeana. Dec. 2014. URL: http://pro.europeana.eu/page/edm-documentation.

[47] Antoine Isaac and Bernhard Haslhofer. "Europeana Linked Open Data – data.europeana.eu." In: *Semantic web journal* 4.3 (Jan. 2013), pp. 291–297. DOI: 10.3233/SW-120092.

[48] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman II. "Five stars of Linked Data vocabulary use." In: *Semantic web journal* 5.3 (2014), pp. 173–176. DOI: 10.3233/SW-140135.

[49] Lizzy Jongma and Chris Dijkshoorn. "Accurator: enriching collections with expert knowledge from the crowd." In: *Proceedings of the museums and the web conference*. Feb. 2016. URL: http://mw2016.museumsandtheweb.com/paper/accurator-enriching-collections-with-expert-knowledge-from-the-crowd/.

[50] Craig A. Knoblock, Pedro Szekely, Eleanor Fink, Duane Degler, David Newbury, Robert Sanderson, Kate Blanch, Sara Snyder, Nilay Chheda, Nimesh Jain, Ravi Raju Krishna, Nikhila Begur Sreekanth, and Yixiang Yao. "Lessons learned in building Linked Data for the American Art Collaborative." In: *Proceedings of the 16th international semantic web conference*. (Vienna, Austria). Ed. by Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin. ISWC '17. Springer International Publishing, 2017, pp. 263–279. DOI: 10.1007/978-3-319-68204-4_26.

[51] Anand Kulkarni, Prayag Narula, David Rolnitzky, and Nathan Kontny. "Wish: amplifying creative ability with expert crowds." In: *Proceedings of the 2nd aaai conference on human computation and crowdsourcing*. (Pittsburgh, PA, USA). HCOMP '14. The AAAI Press, Nov. 2014, pp. 112–120.

[52] Carl Lagoze and Jane Hunter. "The ABC ontology and model." In: *Journal of digital information* 2.2 (2006). ISSN: 1368-7506.

[53] Fuyuko Matsumura, Iwao Kobayashi, Fumihiro Kato, Tetsuro Kamura, Ikki Ohmukai, and Hideaki Takeda. "Producing and consuming linked open data on art with a local community." In: *Proceedings of the 3rd international workshop on consuming linked data*. (Boston, MA, USA). Ed. by Juan Sequeda, Andreas Harth,

and Olaf Hartig. Vol. 905. CEUR Workshop Proceedings. CEUR-WS.org, 2012.

[54] Dmitry Mouromtsev, Peter Haase, Eugene Cherny, Dmitry Pavlov, Alexey Andreev, and Anna Spiridonova. "Towards the Russian linked culture cloud: data enrichment and publishing." In: *Proceedings of the 12th extended semantic web conference*. (Portoroz, Slovenia). Ed. by Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann. ESWC '15. Springer International Publishing, May 2015, pp. 637–651. DOI: 10.1007/978-3-319-18818-8_39.

[55] Julia Noordegraaf, Angela Bartholomew, and Alexandra Eveleigh. "Modeling crowdsourcing for cultural heritage." In: *Proceedings of the museums and the web conference*. Feb. 2014. URL: http://mw2014.museumsandtheweb.com/paper/modeling-crowdsourcing-for-cultural-heritage/.

[56] Leo Obrst, Michael Gruninger, Ken Baclawski, Mike Bennett, Dan Brickley, Gary Berg-Cross, Pascal Hitzler, Krzysztof Janowicz, Christine Kapp, Oliver Kutz, et al. "Semantic web and big data meets applied ontology." In: *Applied ontology* 9.2 (2014), pp. 155–170. DOI: 10.3233/AO-140135.

[57] Johan Oomen and Lora Aroyo. "Crowdsourcing in the cultural heritage domain: opportunities and challenges." In: *Proceedings of the 5th international conference on communities and technologies*. (Brisbane, Australia). C&T '11. ACM, May 2011, pp. 138–149. DOI: 10.1145/2103354.2103373.

[58] Jasper Oosterman and Geert-Jan Houben. "On the invitation of expert contributors from online communities for knowledge crowdsourcing tasks." In: *Proceedings of the 16th international conference on web engineering*. (Lugano, Switzerland). Ed. by Alessandro Bozzon, Philippe Cudre-Maroux, and Cesare Pautasso. ICWE16. Springer, May 2016, pp. 413–421. DOI: 10.1007/978-3-319-38791-8_27.

[59] Alexandre Passant. "Seevl: mining music connections to bring context, search and discovery to the music you like." In: *Semantic web challenge 2011* (2011).

[60] Joris Pekel. *Democratising the rijksmuseum. why did the rijksmuseum make available their highest quality material without restrictions, and what are the results?* Tech. rep. Europeana, 2014.

[61] Alexander J. Quinn and Benjamin B. Bederson. "Human computation: a survey and taxonomy of a growing field." In: *Proceedings of the sigchi conference on human factors in computing systems*. (Vancouver, BC, Canada). CHI '11. ACM, May 2011, pp. 1403–1412. DOI: 10.1145/1978942.1979148.

[62] M. Jordan Raddick, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. "Galaxy zoo: exploring the motivations of citizen science volunteers." In: *Astronomy education review* 9.1 (2010), pp. 1–18. ISSN: 1539-1515. DOI: 10.3847/AER2009036.

[63] Yves Raimond and Tristan Ferne. "The BBC world service archive prototype." In: *Web semantics: science, services and agents on the world wide web* 27-28 (2014). Semantic Web Challenge 2013, pp. 2–9. ISSN: 1570-8268. DOI: 10.1016/j.websem.2014.07.005.

[64] Mia Ridge. "From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing." In: *Curator: the museum journal* 56.4 (2013), pp. 435–450. ISSN: 2151-6952. DOI: 10.1111/cura.12046.

[65] Laurens Rietveld, Wouter Beek, and Stefan Schlobach. "Lod lab: experiments at lod scale." In: *Proceedings of the 14th international semantic web conference*. (Bethlehem, Pennsylvania, USA). Ed. by Marcelo Arenas, Oscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab. Vol. 9367. ISWC '15. Springer International Publishing, 2015, pp. 339–355. ISBN: 978-3-319-25010-6. DOI: https://doi.org/10.1007/978-3-319-25010-6_23.

[66] Cristina Sarasua, Elena Simperl, Natasha Noy, Abraham Bernstein, and Jan Marco Leimeister. "Crowdsourcing and the semantic web: a research manifesto." In: *Human computation journal* 2.1 (2015), pp. 3–17. ISSN: 2330-8001. DOI: 10.15346/hc.v2i1.2.

[67] Guus Schreiber, Alia Amin, Mark Van Assem, Victor De Boer, Lynda Hardman, Michiel Hildebrand, Laura Hollink, Zhisheng Huang, Janneke van Kersen, Marco de Niet, Borys Omelayenko, Jacco van Ossenbruggen, Ronny Siebes, Jos Taekema, Jan Wielemaker, and Bob Wielinge. "MultimediaN E-Culture demonstrator." In: *Proceedings of the 5th international semantic web conference*. (Athens, GA, USA). Ed. by Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora M. Aroyo. Vol. 4273. ISWC '06. Springer Berlin Heidelberg, Nov. 2006, pp. 951–958. DOI: 10.1007/11926078_70.

[68] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. "The semantic web revisited." In: *Ieee intelligent systems* 21.3 (May 2006), pp. 96–101. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.62.

[69] Rainer Simon, Bernhard Haslhofer, Werner Robitza, and Elaheh Momeni Roochi. "Semantically augmented annotations in digitized map collections." In: *Proceedings of the 11th annual international acm/ieee joint conference on digital libraries*. (Ottawa, Ontario,

Canada). JCDL '11. ACM, 2011, pp. 199–202. DOI: 10.1145/1998076.1998114.

[70] Els Stronks. *Negotiating differences: word, image and religion in the dutch republic*. Ed. by Andrew Colin Gow. Vol. 155. Brill, 2011.

[71] Rudi Studer, Richard Benjamins, and Dieter Fensel. "Knowledge engineering: principles and methods." In: *Data & knowledge engineering* 25.1 (1998), pp. 161–197. ISSN: 0169-023X. DOI: 10.1016/S0169-023X(97)00056-6.

[72] Pedro Szekely, Craig A. Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E. Fink, Rachel Allen, and Georgina Goodlander. "Connecting the smithsonian american art museum to the linked data cloud." In: *Proceedings of the 10th extended semantic web conference*. (Montpellier, France). Ed. by Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph. Vol. 7882. ESWC '13. Springer Berlin Heidelberg, May 2013, pp. 593–607. DOI: 10.1007/978-3-642-38288-8_40.

[73] Anna Tordai, Jacco van Ossenbruggen, Guus Schreiber, and Bob Wielinga. "Aligning large SKOS-like vocabularies: two case studies." In: *Proceedings of the 7th extended semantic web conference*. (Heraklion, Crete, Greece). Ed. by Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache. ESWC '10. Springer Berlin Heidelberg, May 2010, pp. 198–212. DOI: 10.1007/978-3-642-13486-9_14.

[74] Jennifer Trant, Lee Ave, and Bruce Wyman. "Investigating social tagging and folksonomy in art museums with steve.museum." In: *Proceedings of the collaborative web tagging workshop*. (Edinburgh, Scotland). 2006, pp. 1–6.

[75] Myriam C. Traub, Jacco van Ossenbruggen, Jiyin He, and Lynda Hardman. "Measuring the effectiveness of gamesourcing expert oil painting annotations." In: *Proceedings of the 36th european conference on ir research*. (Amsterdam). Ed. by Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann. ECIR '14. Springer, Apr. 2014, pp. 112–123. DOI: 10.1007/978-3-319-06028-6_10.

[76] Charles Valentine and Antoine Isaac. "Enhancing the Europeana Data Model (EDM)." In: *Whitepaper* (May 2015).

[77] Marjolein Van Gendt, Antoine Isaac, Lourens Van Der Meij, and Stefan Schlobach. "Semantic web techniques for multiple views on heterogeneous collections: a case study." In: *Proceedings of the 10th european conference on research and advanced technology for digital libraries*. (Alicante, Spain). Ed. by Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, and Rafael C. Carrasco.

ECDL '06. Springer Berlin Heidelberg, Sept. 2006, pp. 426–437. DOI: 10.1007/11863878_36.

[78] Harry Verwayen, Martijn Arnoldus, and Peter B Kaufman. "The problem of the yellow milkmaid: a business model perspective on open metadata." In: *Whitepaper* 2 (2011).

[79] Denny Vrandečić and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase." In: *Communications of the acm* 57.10 (Sept. 2014), pp. 78–85. DOI: 10.1145/2629489.

[80] Shenghui Wang, Antoine Isaac, Valentine Charles, Rob Koopman, Anthi Agoropoulou, and Titia van der Werf. "Hierarchical structuring of cultural heritage objects within large aggregations." In: *Proceedings of the international conference on theory and practice of digital libraries*. (Valletta, Malta). Ed. by Trond Aalberg, Christos Papatheodorou, Milena Dobreva, Giannis Tsakonas, and Charles J. Farrugia. TPDL '13. Springer Berlin Heidelberg, 2013, pp. 247–259. DOI: 10.1007/978-3-642-40501-3_25.

[81] Yiwen Wang, Natalia Stash, Lora Aroyo, Peter Gorgels, Lloyd Rutledge, and Guus Schreiber. "Recommendations based on semantically enriched museum collections." In: *Web semantics: science, services and agents on the world wide web* 6.4 (2008), pp. 283–290. DOI: 10.1016/j.websem.2008.09.002.

[82] Etienne Wenger, Richard Arnold McDermott, and William Snyder. *Cultivating communities of practice: a guide to managing knowledge*. Harvard Business Press, 2002.

[83] Jan Wielemaker, Michiel Hildebrand, and Jacco Van Ossenbruggen. "Using Prolog as the fundament for applications on the semantic web." In: *Proceedings of the 2nd workshop on applications of logic programming and to the web, semantic web and semantic web services*. (Porto, Portugal). Ed. by S Heymans, A Polleres, E Ruckhaus, D Pearse, and G Gupta. CEUR Workshop Proceedings. CEUR-WS.org, 2007, pp. 84–98.

[84] Jan Wielemaker, Michiel Hildebrand, Jacco van Ossenbruggen, and Guus Schreiber. "Thesaurus-based search in large heterogeneous collections." In: *Proceedings of the 7th international semantic web conference*. (Karlsruhe, Germany). Ed. by Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin, and Krishnaprasad Thirunarayan. ISWC '08. Springer, 2008, pp. 695–708. DOI: 10.1007/978-3-540-88564-1_44.

[85] Jan Wielemaker, Wouter Beek, Michiel Hildebrand, and Jacco van Ossenbruggen. "ClioPatria: a SWI-Prolog infrastructure for the Semantic Web." In: *Semantic web journal* 7.5 (2016), pp. 529–541. DOI: 10.3233/SW-150191.

[86] Bob Wielinga, Guus Schreiber, Jan Wielemaker, and Jac Sandberg. "From thesaurus to ontology." In: *Proceedings of the 1st international conference on knowledge capture*. (Victoria, British Columbia, Canada). K-CAP '01. ACM. 2001, pp. 194–201. DOI: 10.1145/500737.500767.

[87] Poonam Yadav and John Darlington. "Design guidelines for the user-centred collaborative citizen science platforms." In: *Human computation journal* 3.1 (2016), pp. 205–211. ISSN: 2330-8001. DOI: 10.15346/hc.v3i1.11.

[88] Victor de Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. "Nichesourcing: harnessing the power of crowds of experts." In: *Proceedings of the 18th international conference on knowledge engineering and knowledge management*. (Galway, Ireland). Ed. by Annette ten Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Acquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez. EKAW '12. Springer Berlin Heidelberg, Oct. 2012, pp. 16–20. DOI: 10.1007/978-3-642-33876-2_3.

[90] Victor de Boer, Jan Wielemaker, Judith van Gent, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. "Supporting Linked Data production for cultural heritage institutes: the Amsterdam museum case study." In: *Proceedings of the 9th extended semantic web conference*. Ed. by Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti. Vol. 7295. ESWC '12. Springer Berlin Heidelberg, 2012, pp. 733–747. DOI: 10.1007/978-3-642-30284-8_56.

SUMMARY

Semantic Web technologies allow cultural heritage institutions to publish interconnected, interoperable data, with explicit semantics. The source of the data published by museums is often the basic metadata recorded in systems aimed at collection management. As a consequence, users are deprived of the curated contextual information of regular exhibitions. To address this problem and provide better access to online collections, institutions employ various approaches to improve object descriptions. Among them is crowdsourcing, a quick and inexpensive source of large quantities of descriptions. However, it remains a challenge to ensure the quality of crowdsourced information, especially for knowledge-intensive tasks. In this thesis, we introduce *nichesourcing*, a method to solve knowledge-intensive tasks, by identifying and engaging small groups of experts. We present a five-step method, to enrich and contextualize object metadata using nichesourcing, thereby improving access to online cultural heritage collections.

ANALYSIS OF COLLECTION DATA The first step of the method concerns the analysis of collection data. During this step, we assess the suitability of the chosen data model and the number of references to external datasets. In Chapter 2, we analyze the Linked Data of the Rijksmuseum Amsterdam, as a case study. The Rijksmuseum collection comprises over a million objects, of which only a fraction can be on display at a given time. To open up the remaining collection, the museum started to digitize objects and publish the resulting information online. The Linked Data of the museum consists of over 22 million statements, describing over 350,000 objects, of which more than 207,000 include a reference to an image. The data is used to support search, recommendation, collection integration and browsing.

The Rijksmuseum uses contextual concepts from structured vocabularies to describe objects. While the museum maintains its own vocabularies to preserve its own perspective, an increasing number of contextual concepts is related to external datasets. The collection data is structured using the Europeana Data Model. Not all aspects of the collection can be captured with the modeling constructs recommended by Europeana. Therefore, we discuss modeling challenges and proposed solutions for contextualizing cultural heritage data in the next chapter.

CONTEXTUALIZATION OF CULTURAL HERITAGE DATA Ontologies make the semantics of data explicit, by providing a shared con-

ceptualization. When a cultural heritage institution wants to publish Linked Data, it is confronted with the choice of which ontology to use. This decision has implications for the source data that can be included, as well as the structure of the resulting Linked Data. As part of the five-step method, we focus in Chapter 3 on how ontologies can be used to structure and represent contextual information about objects in cultural heritage collections. We discuss modeling challenges that regard specialization, object- and event-centric approaches, temporality, representations, views and subject matter. For each challenge, we show modeling approaches of two ontologies often used in the cultural heritage domain: the Europeana Data Model and the CIDOC Conceptual Reference Model.

Based on the discussed modeling challenges, we formulate six requirements for cultural heritage ontologies: 1) the ability to specialize an ontology without decreasing its interoperability, 2) support for recording both attributes as well as events related to objects, 3) ability to capture changes over time, 4) ability to separate descriptions of objects and their representations, 5) support for capturing multiple sources describing the same object and 6) possibility to contextualize objects using subject matter. By considering these requirements, institutions can make a more informed choice when deciding on which ontology to use to contextualize data published online.

NICHESOURCING The usefulness of cultural heritage data hinges on the quality and diversity of descriptions of collection objects. In many cases, existing descriptions are insufficient for retrieval and research tasks, resulting in the need for additional annotations. Eliciting such annotations is a challenge, since it often requires domain-specific knowledge. Where crowdsourcing can be successfully used to execute simple annotation tasks, identifying people with the required expertise might prove challenging for more complex and domain-specific tasks. Nichesourcing addresses this problem, by tapping into the expert knowledge available in niche communities.

In Chapter 4, we present Accurator, a methodology for conducting nichesourcing campaigns, by addressing communities, organizing events and tailoring a web-based annotation tool to a domain of choice. The contributions are the following: 1) a nichesourcing methodology, 2) an annotation tool for experts, 3) validation of the methodology in three case studies and 4) a dataset including the obtained annotations. The case studies concern birds on art, bible prints and fashion images. We compare the quality and quantity of obtained annotations, showing that the nichesourcing methodology in combination with the image annotation tool can be used to collect high-quality annotations in a variety of domains. A user evaluation indicates the tool is suited and usable for domain-specific annotation tasks.

DIVERSIFICATION OF SEARCH RESULTS In Chapter 5, we consider whether, and to what extent, additional semantics in the form of Linked Data can support explorative search. As a case study, we use the Linked Data of the Rijksmuseum, extended with various structured vocabularies. We apply an existing graph search algorithm to this data, which finds paths in the graph from the search term to target objects. Next, the algorithm clusters results with similar paths together. We use the number of resulting clusters and the path length as indicators of diversity. As sample queries, we collected the terms in the museum's query log for the duration of one month.

The results show that for this application domain, the added semantics lead to 1) an increase in the number of results, and 2) an increase in the variety of search results. We hypothesize that the following two factors impact the usefulness of vocabularies for search: 1) the number of links between distinct concepts and the metadata objects and 2) the richness of the internal links between concepts in vocabularies. This fourth step of the method illustrates that additional semantics provided by structured vocabularies can help users to explore collections and reach more objects related to their interest.

INTEGRATION OF COLLECTIONS Online cultural heritage collections often contain complementary objects, which makes integration of heterogeneous collections a worthwhile effort. In Chapter 6, we present the DigiBird system that provides access to four distinct nature-related collections and reinforces crowdsourcing initiatives. The system is designed to harmonize complementary collection objects, make crowd contributions instantaneously available and allow the monitoring of multiple crowdsourcing systems using one dashboard.

Harmonizing data from multiple systems that adhere to different standards proved to be a challenge. The data originates from dynamic systems, as a continuous stream of crowd contributions alters and extends the datasets. With the DigiBird system, institutions can decide to use data in an early stage of the crowdsourcing process. Additionally, undertaking crowdsourcing projects together allows the sharing of resources and provides insights into the time needed to collect results. By leveraging standardized data models and annotations from structured vocabularies, the DigiBird system illustrates the added value of enrichments and the benefits of Linked Data for collection integration.

CONCLUSION In this thesis, we present a method to contextualize and enrich cultural heritage collections, to support explorative search and collection integration. Museums with similar collections as the Rijksmuseum will be able to use the method without requiring major changes. We expect that many steps of the method can be utilized in other domains as well. Disseminating high-quality information about

objects is embedded in the mission of cultural heritage institutions. Institutions that want to publish rich contextualized data online, have to assess the quality of external structured vocabularies and find efficient approaches to relate collection data to these new sources. Nichesourcing is one of these approaches, additionally providing ways to engage with the public. To keep contributors motivated, it is essential that they know their work matters. We show the direct impact of annotations on search functionality and collection integration, highlighting the potential of crowd contributions and Linked Data.

SAMENVATTING

Semantic Web technologie stelt cultureel-erfgoedinstellingen in staat om data te publiceren, waarvan de betekenis is vastgelegd en welke van context is voorzien door middel van externe bronnen. Beperkte metadata over objecten, bedoeld voor het beheren van collecties, vormt veelal de basis voor de gepubliceerde data. De uitgebreide informatie waar tentoonstellingen normaal gezegd in voorzien, ontbreekt daardoor voor gebruikers. Om dit probleem aan te pakken en de toegang tot online collecties te verbeteren, zijn instellingen begonnen met het verbeteren van objectbeschrijvingen. *Crowdsourcing* is een manier om snel grote hoeveelheden beschrijvingen te verzamelen. Het blijft echter een uitdaging om de kwaliteit te borgen van informatie die door crowdsourcing verkregen is. Dit geldt in het bijzonder voor annotatietaken die specifieke kennis vereisen. In deze dissertatie introduceren we *nichesourcing*, een methode voor het uitvoeren van kennisintensieve taken, waarbij gespecialiseerde groepen in het annotatieproces worden betrokken. Deze dissertatie beschrijft een methode om met behulp van nichesourcing objectbeschrijvingen te verrijken en te contextualiseren, waardoor online collecties beter toegankelijk worden. Deze methode bestaat uit vijf stappen.

ANALYSE VAN COLLECTIEDATA De eerste stap van de methode betreft de analyse van collectiedata. We kijken tijdens deze stap naar de geschiktheid van het gekozen datamodel en het aantal verwijzingen naar externe datasets. In hoofdstuk 2 analyseren we als casestudy de *Linked Data* van het Rijksmuseum Amsterdam. De Rijksmuseum collectie bestaat uit meer dan een miljoen objecten, waar op enig moment enkel een fractie kan worden tentoongesteld. Om ook de rest van de collectie toegankelijk te maken, is het museum gestart met het digitaliseren van objecten en het online publiceren van informatie. De Linked Data van het museum bestaat uit 2.846.996 *statements*, die 351.814 objecten beschrijven, waarvan 207.441 een corresponderende afbeelding hebben. De data wordt gebruikt om de collectie toegankelijk en doorzoekbaar te maken, relevante objecten aan te raden en de collectie te integreren met andere collecties.

Het Rijksmuseum gebruikt concepten van gestructureerde vocabulaires om objecten te beschrijven. Ondanks dat het museum er ook voor kiest een eigen vocabulaire te onderhouden, wordt een toenemend aantal concepten gerelateerd aan externe datasets. De collectiedata wordt gestructureerd met behulp van het Europeana Data Model. Niet alle aspecten van de objecten kunnen echter adequaat worden beschreven met de door Europeana voorgeschreven elementen.

Daarom bespreken we de uitdagingen van het modelleren en contextualiseren van cultureel-erfgoeddata in het volgende hoofdstuk.

CULTUREEL-ERFGOEDDATA CONTEXTUALISEREN Een ontologie maakt de betekenis van data expliciet, door middel van een gedeelde conceptualisatie. Wanneer een cultureel-erfgoedinstelling Linked Data wil publiceren, moet er worden gekozen welke ontologie het meest geschikt is. Deze beslissing heeft gevolgen voor welke data er kan worden opgenomen in de dataset, maar ook voor de structuur van de resulterende Linked Data. In hoofdstuk 3 kijken we hoe ontologiën gebruikt kunnen worden voor het structureren en representeren van contextuele informatie over objecten in cultureel-erfgoedcollecties. We bespreken uitdagingen op het gebied van data modelleren, met betrekking tot specialisatie, object- of gebeurtenisgerichte aanpakken, tijd, representatie, perspectieven en onderwerpontsluiting. Elke uitdaging illustreren we met de benaderingen van twee veel gebruikte ontologiën in het cultureel-erfgoeddomein: het Europeana Data Model en het CIDOC Conceptual Reference Model.

Gebaseerd op bovenstaande uitdagingen, formuleren we zes vereisten voor cultureel-erfgoedontologiën: 1) een ontologie kan worden gespecialiseerd zonder dat de interoperabiliteit daar onder lijdt, 2) zowel eigenschappen van objecten, als wel gebeurtenissen gerelateerd aan objecten kunnen worden vastgelegd, 3) veranderingen als gevolg van het verstrijken van tijd kunnen worden beschreven, 4) er kan onderscheid worden gemaakt tussen objecten en hun representaties, 5) verschillende bronnen over hetzelfde object kunnen worden vastgelegd en 6) objecten kunnen van context worden voorzien door middel van onderwerpontsluiting. Instellingen kunnen een afgewogen keuze maken over welke ontologie te gebruiken, wanneer ze deze vereisten in overweging nemen.

NICHESOURCING De waarde van cultureel-erfgoeddata hangt af van de kwaliteit en diversiteit van de beschrijvingen van collectieobjecten. In veel gevallen zijn al bestaande beschrijvingen niet geschikt voor het ondersteunen van onderzoek of voor het online toegankelijk maken van de collectie, waardoor extra annotaties nodig zijn. Het uitvoeren van deze annotatietaken wordt bemoeilijkt doordat er vaak domein-specifieke kennis voor nodig is. Waar crowdsourcing vaak succesvol kan worden gebruikt voor het uitvoeren van eenvoudige taken, is het vinden van mensen met de vereiste expertise voor het uitvoeren van complexere taken vaak een uitdaging. Nichesourcing lost dit probleem op door de expertise van bestaande groepen aan te spreken.

In hoofdstuk 4 presenteren we Accurator: een methode voor het uitvoeren van nichesourcing-campagnes, waarbij specifieke groepen betrokken worden, evenementen worden georganiseerd en een online

applicatie wordt gebruikt, aangepast aan het gekozen domein. Onze bijdragen zijn: 1) een nichesourcing methodologie, 2) een annotatie-applicatie voor experts, 3) validatie van de methodologie in drie case-studies en 4) een dataset met de verzamelde annotaties. De casestudies betreffen kunstwerken met vogelafbeeldingen, bijbelprenten en afbeeldingen van mode. We vergelijken de kwaliteit en kwantiteit van de verkregen annotaties en tonen daarmee aan dat de nichesourcing methodologie in combinatie met de annotatie-applicatie gebruikt kan worden voor het verzamelen van annotaties van hoge kwaliteit in verschillende domeinen. Een gebruikersstudie toont aan dat de applicatie geschikt is voor domein-specifieke annotatietaken.

DIVERSIFICATIE VAN ZOEKRESULTATEN In hoofdstuk 5 onderzoeken we of, en in welke mate, de verrijking van objectbeschrijvingen in de vorm van Linked Data, exploratief zoeken mogelijk maakt. Als casestudy gebruiken we de Linked Data van het Rijksmuseum, gerelateerd aan verschillende gestructureerde vocabulaires. We passen een bestaand zoekalgoritme voor graafstructuren toe op de data, welke verbanden vindt tussen een zoekterm en objecten. Daarna groepeert het algoritme soortgelijke objecten gebaseerd op de gevonden verbanden. We gebruiken het aantal gevonden verbanden en de afstand tussen zoekterm en object als indicator voor diversiteit. De zoektermen die we gebruiken zijn in de loop van een maand ingevoerd door gebruikers van de website van het museum.

De resultaten tonen aan dat binnen dit domein de verrijking leidt tot 1) een toename van het aantal zoekresultaten en 2) een toename van de variatie in zoekresultaten. Onze hypothese is dat er twee factoren invloed hebben op de geschiktheid van vocabulaires voor exploratief zoeken: 1) het aantal connecties tussen unieke concepten en objecten en 2) de rijkdom van interne connecties tussen concepten in vocabulaires. Deze vierde stap van de methode illustreert dat toegevoegde betekenis in de vorm van gestructureerde vocabulaires gebruikers kan helpen bij het verkennen van collecties en het bereiken van objecten waarin zij geïnteresseerd zijn.

INTEGRATIE VAN COLLECTIES Online collecties van verschillende cultureel-erfgoedinstellingen bevatten vaak objecten die elkaar aanvullen, wat het integreren van heterogene collecties de moeite waard maakt. In hoofdstuk 6 beschrijven we het DigiBird-systeem. Dit systeem geeft gebruikers toegang tot vier verschillende collecties en vereenvoudigt crowdsourcing initiatieven. Het systeem is ontworpen om metadata van collectieobjecten op elkaar af te stemmen, crowdsourcing bijdragen onmiddellijk beschikbaar te maken en de voortgang van crowdsourcing initiatieven in de gaten te houden op één centrale plek.

Het is uitdagend om data te harmoniseren van verschillende systemen, welke gebruik maken van een veelvoud aan standaarden en protocollen. De onderliggende data is onderhevig aan een continue stroom van crowdsourcing bijdragen. Met het DigiBird-systeem kunnen instellingen beslissen om data in een vroeg stadium van het crowdsourcing proces te gebruiken. Door het samen uitvoeren van crowdsourcing initiatieven, kunnen benodigde middelen worden gedeeld en kan er inzicht worden verkregen in de tijd die nodig is om goede resultaten te behalen. Met het gebruik van gestandaardiseerde datamodellen en annotaties uit gestructureerde vocabulaires, illustreert het DigiBird-systeem de toegevoegde waarde van verrijkingen en de voordelen van Linked Data voor collectie integratie.

CONCLUSIE In deze dissertatie presenteren we een methode voor het contextualiseren en verrijken van cultureel-erfgoedcollecties, om daarmee exploratief zoeken en collectie-integratie mogelijk te maken. Musea met soortgelijke collecties als het Rijksmuseum zullen de methode kunnen toepassen zonder dat grote aanpassingen nodig zijn. We verwachten dat veel van de stappen van deze methode ook in andere domeinen toepasbaar zijn. Het uitdragen van kwalitatieve informatie over objecten is onderdeel van de missie van cultureel-erfgoedinstellingen. Instellingen die rijke, gecontextualiseerde data online willen publiceren, moeten een inschatting maken van de kwaliteit van beschikbare externe vocabulaires en efficiënte methodes vinden om collectiedata te relateren aan deze nieuwe bronnen. Nichesourcing is één van deze methodes, welke instellingen in staat stelt om op een nieuwe manier het publiek te betrekken. Om mensen te motiveren is het van belang dat ze weten dat hun werk er toe doet. In onze systemen laten wij direct de invloed zien van annotaties op zoekfunctionaliteit en de mogelijkheid tot collectie integratie, waarmee we de kracht van crowdsourcing en Linked Data benadrukken.

SIKS DISSERTATION SERIES

2011

2011-1 Botond Cseke (RUN) Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-2 Nick Tinnemeier (UU) Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

2011-3 Jan Martijn van der Werf (TUE) Compositional Design and Verification of Component-Based Information Systems

2011-4 Hado van Hasselt (UU) Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms

2011-5 Base van der Raadt (VU) Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

2011-6 Yiwen Wang (TUE) Semantically-Enhanced Recommendations in Cultural Heritage

2011-7 Yujia Cao (UT) Multimodal Information Presentation for High Load Human Computer Interaction

2011-8 Nieske Vergunst (UU) BDI-based Generation of Robust Task-Oriented Dialogues

2011-9 Tim de Jong (OU) Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT) Cloud Content Contention

2011-11 Dhaval Vyas (UT) Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE) Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT) Airport under Control. Multi-agent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR) Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA) The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM) Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA) Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18 Mark Ponsen (UM) Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU) The Mind ' s Eye on Personal Profiles

2011-20 Qing Gu (VU) Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD) Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA) System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA) Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT) Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU) Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU) Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU) Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein (UVA) Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-29 Faisal Kamiran (TUE) Discrimination-aware Classification

2011-30 Egon van den Broek (UT) Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-31 Ludo Waltman (EUR) Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-32 Nees-Jan van Eck (EUR) Methodological Advances in Bibliometric Mapping of Science

2011-33 Tom van der Weide (UU) Arguing to Motivate Decisions

2011-34 Paolo Turrini (UU) Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-35 Maaike Harbers (UU) Explaining Agent Behavior in Virtual Training

2011-36 Erik van der Spek (UU) Experiments in serious game design: a cognitive approach

2011-37 Adriana Burlutiu (RUN) Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

2011-38 Nyree Lemmens (UM) Bee-inspired Distributed Optimization

2011-39 Joost Westra (UU) Organizing Adaptation using Agents in Serious Games

2011-40 Viktor Clerc (VU) Architectural Knowledge Management in Global Software Development
2011-41 Luan Ibraimi (UT) Cryptographically Enforced Distributed Data Access Control
2011-42 Michal Sindlar (UU) Explaining Behavior through Mental State Attribution
2011-43 Henk van der Schuur (UU) Process Improvement through Software Operation Knowledge
2011-44 Boris Reuderink (UT) Robust Brain-Computer Interfaces
2011-45 Herman Stehouwer (UvT) Statistical Language Models for Alternative Sequence Selection
2011-46 Beibei Hu (TUD) Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
2011-47 Azizi Bin Ab Aziz (VU) Exploring Computational Models for Intelligent Support of Persons with Depression
2011-48 Mark Ter Maat (UT) Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
2011-49 Andreea Niculescu (UT) Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

2012-1 Terry Kakeeto (UvT) Relationship Marketing for SMEs in Uganda
2012-2 Muhammad Umair(VU) Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
2012-3 Adam Vanya (VU) Supporting Architecture Evolution by Mining Software Repositories
2012-4 Jurriaan Souer (UU) Development of Content Management System-based Web Applications
2012-5 Marijn Plomp (UU) Maturing Interorganisational Information Systems
2012-6 Wolfgang Reinhardt (OU) Awareness Support for Knowledge Workers in Research Networks
2012-7 Rianne van Lambalgen (VU) When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
2012-8 Gerben de Vries (UVA) Kernel Methods for Vessel Trajectories
2012-9 Ricardo Neisse (UT) Trust and Privacy Management Support for Context-Aware Service Platforms
2012-10 David Smits (TUE) Towards a Generic Distributed Adaptive Hypermedia Environment
2012-11 J.C.B. Rantham Prabhakara (TUE) Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
2012-12 Kees van der Sluijs (TUE) Model Driven Design and Data Integration in Semantic Web Information Systems
2012-13 Suleman Shahid (UvT) Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
2012-14 Evgeny Knutov (TUE) Generic Adaptation Framework for Unifying Adaptive Web-based Systems
2012-15 Natalie van der Wal (VU) Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
2012-16 Fiemke Both (VU) Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
2012-17 Amal Elgammal (UvT) Towards a Comprehensive Framework for Business Process Compliance
2012-18 Eltjo Poort (VU) Improving Solution Architecting Practices
2012-19 Helen Schonenberg (TUE) What's Next? Operational Support for Business Process Execution
2012-20 Ali Bahramisharif (RUN) Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
2012-21 Roberto Cornacchia (TUD) Querying Sparse Matrices for Information Retrieval
2012-22 Thijs Vis (UvT) Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
2012-23 Christian Muehl (UT) Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
2012-24 Laurens van der Werff (UT) Evaluation of Noisy Transcripts for Spoken Document Retrieval
2012-25 Silja Eckartz (UT) Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
2012-26 Emile de Maat (UVA) Making Sense of Legal Text
2012-27 Hayrettin Gurkok (UT) Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
2012-28 Nancy Pascall (UvT) Engendering Technology Empowering Women
2012-29 Almer Tigelaar (UT) Peer-to-Peer Information Retrieval
2012-30 Alina Pommeranz (TUD) Designing Human-Centered Systems for Reflective Decision Making
2012-31 Emily Bagarukayo (RUN) A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
2012-32 Wietske Visser (TUD) Qualitative multi-criteria preference representation and reasoning
2012-33 Rory Sie (OUN) Coalitions in Cooperation Networks (COCOON)
2012-34 Pavol Jancura (RUN) Evolutionary analysis in PPI networks and applications
2012-35 Evert Haasdijk (VU) Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

2012-36 Denis Ssebugwawo (RUN) Analysis and Evaluation of Collaborative Modeling Processes
2012-37 Agnes Nakakawa (RUN) A Collaboration Process for Enterprise Architecture Creation
2012-38 Selmar Smit (VU) Parameter Tuning and Scientific Testing in Evolutionary Algorithms
2012-39 Hassan Fatemi (UT) Risk-aware design of value and coordination networks
2012-40 Agus Gunawan (UvT) Information Access for SMEs in Indonesia
2012-41 Sebastian Kelle (OU) Game Design Patterns for Learning
2012-42 Dominique Verpoorten (OU) Reflection Amplifiers in self-regulated Learning
2012-43 Withdrawn
2012-44 Anna Tordai (VU) On Combining Alignment Techniques
2012-45 Benedikt Kratz (UvT) A Model and Language for Business-aware Transactions
2012-46 Simon Carter (UVA) Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
2012-47 Manos Tsagkias (UVA) Mining Social Media: Tracking Content and Predicting Behavior
2012-48 Jorn Bakker (TUE) Handling Abrupt Changes in Evolving Time-series Data
2012-49 Michael Kaisers (UM) Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
2012-50 Steven van Kervel (TUD) Ontologogy driven Enterprise Information Systems Engineering
2012-51 Jeroen de Jong (TUD) Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013

2013-1 Viorel Milea (EUR) News Analytics for Financial Decision Support
2013-2 Erietta Liarou (CWI) MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
2013-3 Szymon Klarman (VU) Reasoning with Contexts in Description Logics
2013-4 Chetan Yadati(TUD) Coordinating autonomous planning and scheduling
2013-5 Dulce Pumareja (UT) Groupware Requirements Evolutions Patterns
2013-6 Romulo Goncalves(CWI) The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
2013-7 Giel van Lankveld (UvT) Quantifying Individual Player Differences
2013-8 Robbert-Jan Merk(VU) Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
2013-9 Fabio Gori (RUN) Metagenomic Data Analysis: Computational Methods and Applications
2013-10 Jeewanie Jayasinghe Arachchige(UvT) A Unified Modeling Framework for Service Design.
2013-11 Evangelos Pournaras(TUD) Multi-level Reconfigurable Self-organization in Overlay Services
2013-12 Marian Razavian(VU) Knowledge-driven Migration to Services
2013-13 Mohammad Safiri(UT) Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
2013-14 Jafar Tanha (UVA) Ensemble Approaches to Semi-Supervised Learning Learning
2013-15 Daniel Hennes (UM) Multiagent Learning - Dynamic Games and Applications
2013-16 Eric Kok (UU) Exploring the practical benefits of argumentation in multi-agent deliberation
2013-17 Koen Kok (VU) The PowerMatcher: Smart Coordination for the Smart Electricity Grid
2013-18 Jeroen Janssens (UvT) Outlier Selection and One-Class Classification
2013-19 Renze Steenhuizen (TUD) Coordinated Multi-Agent Planning and Scheduling
2013-20 Katja Hofmann (UvA) Fast and Reliable Online Learning to Rank for Information Retrieval
2013-21 Sander Wubben (UvT) Text-to-text generation by monolingual machine translation
2013-22 Tom Claassen (RUN) Causal Discovery and Logic
2013-23 Patricio de Alencar Silva(UvT) Value Activity Monitoring
2013-24 Haitham Bou Ammar (UM) Automated Transfer in Reinforcement Learning
2013-25 Agnieszka Anna Latoszek-Berendsen (UM) Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
2013-26 Alireza Zarghami (UT) Architectural Support for Dynamic Homecare Service Provisioning
2013-27 Mohammad Huq (UT) Inference-based Framework Managing Data Provenance
2013-28 Frans van der Sluis (UT) When Complexity becomes Interesting: An Inquiry into the Information eXperience
2013-29 Iwan de Kok (UT) Listening Heads
2013-30 Joyce Nakatumba (TUE) Resource-Aware Business Process Management: Analysis and Support
2013-31 Dinh Khoa Nguyen (UvT) Blueprint Model and Language for Engineering Cloud Applications
2013-32 Kamakshi Rajagopal (OUN) Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
2013-33 Qi Gao (TUD) User Modeling and Personalization in the Microblogging Sphere
2013-34 Kien Tjin-Kam-Jet (UT) Distributed Deep Web Search

2013-35 Abdallah El Ali (UvA) Minimal Mobile Human Computer Interaction
2013-36 Than Lam Hoang (TUE) Pattern Mining in Data Streams
2013-37 Dirk Borner (OUN) Ambient Learning Displays
2013-38 Eelco den Heijer (VU) Autonomous Evolutionary Art
2013-39 Joop de Jong (TUD) A Method for Enterprise Ontology based Design of Enterprise Information Systems
2013-40 Pim Nijssen (UM) Monte-Carlo Tree Search for Multi-Player Games
2013-41 Jochem Liem (UVA) Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
2013-42 Leon Planken (TUD) Algorithms for Simple Temporal Reasoning
2013-43 Marc Bron (UVA) Exploration and Contextualization through Interaction and Concepts

2014

2014-1 Nicola Barile (UU) Studies in Learning Monotone Models from Data
2014-2 Fiona Tuliyano (RUN) Combining System Dynamics with a Domain Modeling Method
2014-3 Sergio Raul Duarte Torres (UT) Information Retrieval for Children: Search Behavior and Solutions
2014-4 Hanna Jochmann-Mannak (UT) Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
2014-5 Jurriaan van Reijsen (UU) Knowledge Perspectives on Advancing Dynamic Capability
2014-6 Damian Tamburri (VU) Supporting Networked Software Development
2014-7 Arya Adriansyah (TUE) Aligning Observed and Modeled Behavior
2014-8 Samur Araujo (TUD) Data Integration over Distributed and Heterogeneous Data Endpoints
2014-9 Philip Jackson (UvT) Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
2014-10 Ivan Salvador Razo Zapata (VU) Service Value Networks
2014-11 Janneke van der Zwaan (TUD) An Empathic Virtual Buddy for Social Support
2014-12 Willem van Willigen (VU) Look Ma, No Hands: Aspects of Autonomous Vehicle Control
2014-13 Arlette van Wissen (VU) Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
2014-14 Yangyang Shi (TUD) Language Models With Meta-information
2014-15 Natalya Mogles (VU) Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
2014-16 Krystyna Milian (VU) Supporting trial recruitment and design by automatically interpreting eligibility criteria
2014-17 Kathrin Dentler (VU) Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
2014-18 Mattijs Ghijsen (UVA) Methods and Models for the Design and Study of Dynamic Agent Organizations
2014-19 Vinicius Ramos (TUE) Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
2014-20 Mena Habib (UT) Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
2014-21 Kassidy Clark (TUD) Negotiation and Monitoring in Open Environments
2014-22 Marieke Peeters (UU) Personalized Educational Games - Developing agent-supported scenario-based training
2014-23 Eleftherios Sidirourgos (UVA/CWI) Space Efficient Indexes for the Big Data Era
2014-24 Davide Ceolin (VU) Trusting Semi-structured Web Data
2014-25 Martijn Lappenschaar (RUN) New network models for the analysis of disease interaction
2014-26 Tim Baarslag (TUD) What to Bid and When to Stop
2014-27 Rui Jorge Almeida (EUR) Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
2014-28 Anna Chmielowiec (VU) Decentralized k-Clique Matching
2014-29 Jaap Kabbedijk (UU) Variability in Multi-Tenant Enterprise Software
2014-30 Peter de Cock (UvT) Anticipating Criminal Behaviour
2014-31 Leo van Moergestel (UU) Agent Technology in Agile Multiparallel Manufacturing and Product Support
2014-32 Naser Ayat (UvA) On Entity Resolution in Probabilistic Data
2014-33 Tesfa Tegegne (RUN) Service Discovery in eHealth
2014-34 Christina Manteli (VU) The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
2014-35 Joost van Ooijen (UU) Cognitive Agents in Virtual Worlds: A Middleware Design Approach
2014-36 Joos Buijs (TUE) Flexible Evolutionary Algorithms for Mining Structured Process Models

2014-37 Maral Dadvar (UT) Experts and Machines United Against Cyberbullying
2014-38 Danny Plass-Oude Bos (UT) Making brain-computer interfaces better: improving usability through post-processing.
2014-39 Jasmina Maric (UvT) Web Communities, Immigration, and Social Capital
2014-40 Walter Omona (RUN) A Framework for Knowledge Management Using ICT in Higher Education
2014-41 Frederic Hogenboom (EUR) Automated Detection of Financial Events in News Text
2014-42 Carsten Eijckhof (CWI/TUD) Contextual Multidimensional Relevance Models
2014-43 Kevin Vlaanderen (UU) Supporting Process Improvement using Method Increments
2014-44 Paulien Meesters (UvT) Intelligent Blauw. Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
2014-45 Birgit Schmitz (OUN) Mobile Games for Learning: A Pattern-Based Approach
2014-46 Ke Tao (TUD) Social Web Data Analytics: Relevance, Redundancy, Diversity
2014-47 Shangsong Liang (UVA) Fusion and Diversification in Information Retrieval

2015

2015-1 Niels Netten (UvA) Machine Learning for Relevance of Information in Crisis Response
2015-2 Faiza Bukhsh (UvT) Smart auditing: Innovative Compliance Checking in Customs Controls
2015-3 Twan van Laarhoven (RUN) Machine learning for network data
2015-4 Howard Spoelstra (OUN) Collaborations in Open Learning Environments
2015-5 Christoph Bosch (UT) Cryptographically Enforced Search Pattern Hiding
2015-6 Farideh Heidari (TUD) Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
2015-7 Maria-Hendrike Peetz (UvA) Time-Aware Online Reputation Analysis
2015-8 Jie Jiang (TUD) Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
2015-9 Randy Klaassen (UT) HCI Perspectives on Behavior Change Support Systems
2015-10 Henry Hermans (OUN) OpenU: design of an integrated system to support lifelong learning
2015-11 Yongming Luo (TUE) Designing algorithms for big graph datasets: A study of computing bisimulation and joins
2015-12 Julie M. Birkholz (VU) Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
2015-13 Giuseppe Procaccianti (VU) Energy-Efficient Software
2015-14 Bart van Straalen (UT) A cognitive approach to modeling bad news conversations
2015-15 Klaas Andries de Graaf (VU) Ontology-based Software Architecture Documentation
2015-16 Changyun Wei (UT) Cognitive Coordination for Cooperative Multi-Robot Teamwork
2015-17 Andre van Cleeff (UT) Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
2015-18 Holger Pirk (CWI) Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
2015-19 Bernardo Tabuenca (OUN) Ubiquitous Technology for Lifelong Learners
2015-20 Lois Vanhee (UU) Using Culture and Values to Support Flexible Coordination
2015-21 Sibren Fetter (OUN) Using Peer-Support to Expand and Stabilize Online Learning
2015-22 Zhemin Zhu (UT) Co-occurrence Rate Networks
2015-23 Luit Gazendam (VU) Cataloguer Support in Cultural Heritage
2015-24 Richard Berendsen (UVA) Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
2015-25 Steven Woudenberg (UU) Bayesian Tools for Early Disease Detection
2015-26 Alexander Hogenboom (EUR) Sentiment Analysis of Text Guided by Semantics and Structure
2015-27 Sandor Heman (CWI) Updating compressed colomn stores
2015-28 Janet Bagorogoza (TiU) Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
2015-29 Hendrik Baier (UM) Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
2015-30 Kiavash Bahreini (OU) Real-time Multimodal Emotion Recognition in E-Learning
2015-31 Yakup Koe (TUD) On the robustness of Power Grids
2015-32 Jerome Gard (UL) Corporate Venture Management in SMEs
2015-33 Frederik Schadd (TUD) Ontology Mapping with Auxiliary Resources
2015-34 Victor de Graaf (UT) Gesocial Recommender Systems
2015-35 Jungxao Xu (TUD) Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

2016-1 Syed Saiden Abbas (RUN) Recognition of Shapes by Humans and Machines

2016-2 Michiel Christiaan Meulendijk (UU) Optimizing medication reviews through decision support: prescribing a better pill to swallow

2016-3 Maya Sappelli (RUN) Knowledge Work in Context: User Centered Knowledge Worker Support

2016-4 Laurens Rietveld (VU) Publishing and Consuming Linked Data

2016-5 Evgeny Sherkhonov (UVA) Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

2016-6 Michel Wilson (TUD) Robust scheduling in an uncertain environment

2016-7 Jeroen de Man (VU) Measuring and modeling negative emotions for virtual training

2016-8 Matje van de Camp (TiU) A Link to the Past: Constructing Historical Social Networks from Unstructured Data

2016-9 Archana Nottamkandath (VU) Trusting Crowdsourced Information on Cultural Artefacts

2016-10 George Karafotias (VUA) Parameter Control for Evolutionary Algorithms

2016-11 Anne Schuth (UVA) Search Engines that Learn from Their Users

2016-12 Max Knobbout (UU) Logics for Modelling and Verifying Normative Multi-Agent Systems

2016-13 Nana Baah Gyan (VU) The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

2016-14 Ravi Khadka (UU) Revisiting Legacy Software System Modernization

2016-15 Steffen Michels (RUN) Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

2016-16 Guangliang Li (UVA) Socially Intelligent Autonomous Agents that Learn from Human Reward

2016-17 Berend Weel (VU) Towards Embodied Evolution of Robot Organisms

2016-18 Albert Merono Penuela (VU) Refining Statistical Data on the Web

2016-19 Julia Efremova (TUE) Mining Social Structures from Genealogical Data

2016-20 Daan Odijk (UVA) Context & Semantics in News & Web Search

2016-21 Alejandro Moreno Colleri (UT) From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

2016-22 Grace Lewis (VU) Software Architecture Strategies for Cyber-Foraging Systems

2016-23 Fei Cai (UVA) Query Auto Completion in Information Retrieval

2016-24 Brend Wanders (UT) Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

2016-25 Julia Kiseleva (TUE) Using Contextual Information to Understand Searching and Browsing Behavior

2016-26 Dilhan Thilakarathne (VU) In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

2016-27 Wen Li (TUD) Understanding Geo-spatial Information on Social Media

2016-28 Mingxin Zhang (TUD) Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

2016-29 Nicolas Honing (TUD) Peak reduction in decentralised electricity systems -Markets and prices for flexible planning

2016-30 Ruud Mattheij (UvT) The Eyes Have It

2016-31 Mohammad Khelghati (UT) Deep web content monitoring

2016-32 Eelco Vriezekolk (UT) Assessing Telecommunication Service Availability Risks for Crisis Organisations

2016-33 Peter Bloem (UVA) Single Sample Statistics, exercises in learning from just one example

2016-34 Dennis Schunselaar (TUE) Configurable Process Trees: Elicitation, Analysis, and Enactment

2016-35 Zhaochun Ren (UVA) Monitoring Social Media: Summarization, Classification and Recommendation

2016-36 Daphne Karreman (UT) Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

2016-37 Giovanni Sileno (UvA) Aligning Law and Action - a conceptual and computational inquiry

2016-38 Andrea Minuto (UT) Materials that Matter - Smart Materials meet Art & Interaction Design

2016-39 Merijn Bruijnes (UT) Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

2016-40 Christian Detweiler (TUD) Accounting for Values in Design

2016-41 Thomas King (TUD) Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

2016-42 Spyros Martzoukos (UVA) Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

2016-43 Saskia Koldijk (RUN) Context-Aware Support for Stress Self-Management: From Theory to Practice

2016-44 Thibault Sellam (UVA) Automatic Assistants for Database Exploration

2016-45 Bram van de Laar (UT) Experiencing Brain-Computer Interface Control

2016-46 Jorge Gallego Perez (UT) Robots to Make you Happy

2016-47 Christina Weber (UL) Real-time foresight - Preparedness for dynamic innovation networks
2016-48 Tanja Buttler (TUD) Collecting Lessons Learned
2016-49 Gleb Polevoy (TUD) Participation and Interaction in Projects. A Game-Theoretic Analysis
2016-50 Yan Wang (UVT) The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

2017-1 Jan-Jaap Oerlemans (UL) Investigating Cybercrime
2017-2 Sjoerd Timmer (UU) Designing and Understanding Forensic Bayesian Networks using Argumentation
2017-3 Daniel Harold Telgen (UU) Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
2017-4 Mrunal Gawade (CWI) Multi-Core Parallelism in a Column-Store
2017-5 Mahdieh Shadi (UVA) Collaboration Behavior
2017-6 Damir Vandic (EUR) Intelligent Information Systems for Web Product Search
2017-7 Roel Bertens (UU) Insight in Information: from Abstract to Anomaly
2017-8 Rob Konijn (VU) Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
2017-9 Dong Nguyen (UT) Text as Social and Cultural Data: A Computational Perspective on Variation in Text
2017-10 Robby van Delden (UT) (Steering) Interactive Play Behavior
2017-11 Florian Kunneman (RUN) Modelling patterns of time and emotion in Twitter #anticipointment
2017-12 Sander Leemans (TUE) Robust Process Mining with Guarantees
2017-13 Gijs Huisman (UT) Social Touch Technology - Extending the reach of social touch through haptic technology
2017-14 Shoshannah Tekofsky (UvT) You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
2017-15 Peter Berck, Radboud University (RUN) Memory-Based Text Correction
2017-16 Aleksandr Chuklin (UVA) Understanding and Modeling Users of Modern Search Engines
2017-17 Daniel Dimov (UL) Crowdsourced Online Dispute Resolution
2017-18 Ridho Reinanda (UVA) Entity Associations for Search
2017-19 Jeroen Vuurens (TUD) Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
2017-20 Mohammadbashir Sedighi (TUD) Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
2017-21 Jeroen Linssen (UT) Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
2017-22 Sara Magliacane (VU) Logics for causal inference under uncertainty
2017-23 David Graus (UVA) Entities of Interest - Discovery in Digital Traces
2017-24 Chang Wang (TUD) Use of Affordances for Efficient Robot Learning
2017-25 Veruska Zamborlini (VU) Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
2017-26 Merel Jung (UT) Socially intelligent robots that understand and respond to human touch
2017-27 Michiel Joosse (UT) Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
2017-28 John Klein (VU) Architecture Practices for Complex Contexts
2017-29 Adel Alhuraibi (UVT) From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT
2017-30 Wilma Latuny (UVT) The Power of Facial Expressions
2017-31 Ben Ruijl (UL) Advances in computational methods for QFT calculations
2017-32 Thaer Samar (RUN) Access to and Retrievability of Content in Web Archives
2017-33 Brigit van Loggem (OU) Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
2017-34 Maren Scheffel (OUN) The Evaluation Framework for Learning Analytics
2017-35 Martine de Vos (VU) Interpreting natural science spreadsheets
2017-36 Yuanhao Guo (UL) Shape Analysis for Phenotype Characterisation from High-throughput Imaging
2017-37 Alejandro Montes Garcia (TUE) WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
2017-38 Alex Kayal (TUD) Normative Social Applications
2017-39 Sara Ahmadi (RUN) Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
2017-40 Altaf Hussain Abro (VUA) Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems"
2017-41 Adnan Manzoor (VUA) Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

2017-42 Elena Sokolova (RUN) Causal discovery from mixed and missing data with applications on ADHD datasets
2017-43 Maaike de Boer (RUN) Semantic Mapping in Video Retrieval
2017-44 Garm Lucassen (UU) Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
2017-45 Bas Testerink (UU) Decentralized Runtime Norm Enforcement
2017-46 Jan Schneider (OU) Sensor-based Learning Support
2017-47 Yie Yang (TUD) Crowd Knowledge Creation Acceleration
2017-48 Angel Suarez (OU) Collaborative inquiry-based learning

2018

2018-1 Han van der Aa (VUA) Comparing and Aligning Process Representations
2018-2 Felix Mannhardt (TUE) Multi-perspective Process Mining
2018-3 Steven Bosems (UT) Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
2018-4 Jordan Janeiro (TUD) Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
2018-5 Hugo Huurdeman (UVA) Supporting the Complex Dynamics of the Information Seeking Process
2018-6 Dan Ionita (UT) Model-Driven Information Security Risk Assessment of Socio-Technical Systems
2018-7 Jieting Luo (UU) A formal account of opportunism in multi-agent systems
2018-8 Rick Smetsers (RUN) Advances in Model Learning for Software Systems
2018-9 Xu Xie (TUD) Data Assimilation in Discrete Event Simulations
2018-10 Julienka Mollee (VUA) Moving forward: supporting physical activity behavior change through intelligent technology
2018-11 Mahdi Sargolzaei (UVA) Enabling Framework for Service-oriented Collaborative Networks
2018-12 Xixi Lu (TUE) Using behavioral context in process mining
2018-13 Seyed Amin Tabatabaei (VUA) Using behavioral context in process mining: Exploring the added value of computational models for increasing the use of renewable energy in the residential sector
2018-14 Bart Joosten (UVT) Detecting Social Signals with Spatiotemporal Gabor Filters
2018-15 Naser Davarzani (UM) Biomarker discovery in heart failure
2018-16 Jaebok Kim (UT) Automatic recognition of engagement and emotion in a group of children
2018-17 Jianpeng Zhang (TUE) On Graph Sample Clustering
2018-18 Henriette Nakad (UL) De Notaris en Private Rechtspraak
2018-19 Minh Duc Pham (VUA) Emergent relational schemas for RDF
2018-20 Manxia Liu (RUN) Time and Bayesian Networks
2018-21 Aad Slootmaker (OUN) EMERGO: a generic platform for authoring and playing scenario-based serious games
2018-22 Eric Fernandes de Mello Araujo (VUA) Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
2018-23 Kim Schouten (EUR) Semantics-driven Aspect-Based Sentiment Analysis
2018-24 Jered Vroon (UT) Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
2018-25 Riste Gligorov (VUA) Serious Games in Audio-Visual Collections
2018-26 Roelof Anne Jelle de Vries (UT) Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
2018-27 Maikel Leemans (TUE) Hierarchical Process Mining for Scalable Software Analysis
2018-28 Christian Willemse (UT) Social Touch Technologies: How they feel and how they make you feel
2018-29 Yu Gu (UVT) Emotion Recognition from Mandarin Speech
2018-30 Wouter Beek (VUA) The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

2019-1 Rob van Eijk (UL) Comparing and Aligning Process Representations
2019-2 Emmanuelle Beauxis Aussalet (CWI, UU) Statistics and Visualizations for Assessing Class Size Uncertainty
2019-3 Eduardo Gonzalez Lopez de Murillas (TUE) Process Mining on Databases: Extracting Event Data from Real Life Data Sources
2019-4 Ridho Rahmadi (RUN) Finding stable causal structures from clinical data
2019-5 Sebastiaan van Zelst (TUE) Process Mining with Streaming Data